

# MARFCAT: Transitioning to Binary and Larger Data Sets of SATE IV

Serguei A. Mokhov<sup>1,2</sup>, Joey Paquet<sup>1</sup>, Mourad Debbabi<sup>1</sup>, Yankui Sun<sup>2</sup>

<sup>1</sup>Concordia University  
Montreal, QC, Canada

{mokhov,paquet,debbabi}@encs.concordia.ca

<sup>2</sup>Tsinghua University  
Beijing, China  
syk@mail.tsinghua.edu.cn

## Abstract

We present a second iteration of a machine learning approach to static code analysis and fingerprinting for weaknesses related to security, software engineering, and others using the open-source MARF framework and the MARFCAT application based on it for the NIST's SATE IV static analysis tool exposition workshop's data sets that include additional test cases, including new large synthetic cases. To aid detection of weak or vulnerable code, including source or binary on different platforms the machine learning approach proved to be fast and accurate to for such tasks where other tools are either much slower or have much smaller recall of known vulnerabilities. We use signal and NLP processing techniques in our approach to accomplish the identification and classification tasks. MARFCAT's design from the beginning in 2010 made is independent of the language being analyzed, source code, bytecode, or binary. In this follow up work with explore some preliminary results in this area. We evaluated also additional algorithms that were used to process the data.

## Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>Related Work</b>	<b>3</b>
<b>3</b>	<b>Data Sets</b>	<b>4</b>
<b>4</b>	<b>Methodology</b>	<b>5</b>
4.1	Methodology Overview . . . . .	5
4.2	CVEs and CWEs – the Knowledge Base . . . . .	6
4.3	Categories for Machine Learning . . . . .	6
4.4	Algorithms . . . . .	7
4.4.1	Signal Pipeline . . . . .	7
4.4.2	NLP Pipeline . . . . .	7
4.5	Binary and Bytecode Analysis . . . . .	7
4.6	Wavelets . . . . .	8
4.7	Demand-Driven Distributed Evaluation with GIPSY . . . . .	9
4.8	Export . . . . .	10
4.8.1	SATE . . . . .	10
4.8.2	Forensic Lucid . . . . .	10
4.8.3	SAFES . . . . .	11
4.9	Experiments . . . . .	11

<b>5</b>	<b>Results</b>	<b>11</b>
5.1	Preliminary Results Summary . . . . .	12
5.2	Version SATE-IV.1 . . . . .	13
5.2.1	Half-Training Data For Training and Full For Testing . . . . .	13
5.3	Version SATE-IV.2 . . . . .	19
5.4	Version SATE-IV.5 . . . . .	20
5.4.1	Wavelet Experiments . . . . .	20
<b>6</b>	<b>Conclusion</b>	<b>21</b>
6.1	Shortcomings . . . . .	21
6.2	Advantages . . . . .	22
6.3	Practical Implications . . . . .	23
6.4	Future Work . . . . .	23
6.5	Acknowledgments . . . . .	24
<b>A</b>	<b>Classification Result Tables</b>	<b>28</b>
<b>B</b>	<b>Forensic Lucid Report Example</b>	<b>28</b>

## List of Tables

1	CVE Stats for Wireshark 1.2.0, Separating DWT Wavelet Filter Preprocessing .	29
2	CVE Stats for Wireshark 1.2.0, Low-Pass FFT Filter Preprocessing . . . . .	30
3	CWE Stats for Wireshark 1.2.0, Separating DWT Wavelet Filter Preprocessing .	31
4	CWE Stats for Wireshark 1.2.0, Low-Pass FFT Filter Preprocessing . . . . .	31

## List of Figures

1	Machine-learning-based static code analysis testing algorithm using the signal pipeline . . . . .	8
2	Machine-learning-based static code analysis testing algorithm using the NLP pipeline . . . . .	9
3	A wave graph of a fraction of the CVE-2009-2562-vulnerable <code>packet-afs.c</code> in Wireshark 1.2.0 . . . . .	14
4	Spectrograms of CVE-2009-2562-vulnerable <code>packet-afs.c</code> in Wireshark 1.2.0, fixed Wireshark 1.2.9 and Wireshark 1.2.18 . . . . .	15
5	A spectrogram of CVE-2009-2562-vulnerable <code>packet-afs.c</code> in Wireshark 1.2.0, after SDWT . . . . .	21

## 1 Introduction

This is a follow up work on the first incarnation of MARFCAT detailed in [Mok10d, Mok11]. Thus, the majority of the results content here addresses the newer iteration duplicating only the necessary background and methodology information (reduced). The reader is deferred to consult the expanded background information and results in that previous work freely accessible online (and the arXiv version of that is still occasionally updated).

We elaborate on the details of the expanded methodology and the corresponding results of application of the machine learning techniques along with signal and NLP processing to static source and binary code analysis in search for weaknesses and vulnerabilities. We use the tool, named *MARFCAT*, a MARF-based Code Analysis Tool [Mok12], first exhibited at the Static Analysis Tool Exposition (SATE) workshop in 2010 [ODBN10] to machine-learn from the (Common Vulnerabilities and Exposures) CVE-based vulnerable as well as synthetic CWE-based cases to verify the fixed versions as well as non-CVE based cases from the projects written in same programming languages. The 2<sup>nd</sup> iteration of this work was prepared for SATE IV [ODBN12] and uses its updated data set and application.

On the NLP side, we employ simple classical NLP techniques ( $n$ -grams and various smoothing algorithms), also combined with machine learning for novel non-NLP applications of detection, classification, and reporting of weaknesses related to vulnerabilities or bad coding practices found in artificial constrained languages, such as programming languages and their compiled counterparts. We compare and contrast the NLP approach to the signal processing approach in our results summary and illustrate concrete results and for the same test cases.

We claim that the presented machine learning approach is novel and highly beneficial in static analysis and routine testing of any kind of code, including source code and binary deployments for its efficiency in terms of speed, relatively high precision, robustness, and being a complimentary tool to other approaches that do in-depth semantic analysis, etc. by prioritizing those tools' targets. All that can be used in automatic manner in distributed and scalable diverse environments to ensure the code safety, especially the mission critical software code in all kinds of systems. It uses spectral, acoustic and language models to learn and classify such a code.

This document, like its predecessor, is a “rolling draft” with several updates expected to be made as the project progresses beyond SATE IV. It is accompanied with the updates to the open-source MARFCAT tool itself [Mok12].

## Organization

The related work, some of the present methodology is based on, is referenced in Section 2. The methodology summary is in Section 4. We present some of the results in Section 5 from the SAMATE reference test data set. Then we present a brief summary, description of the limitations of the current realization of the approach and concluding remarks in Section 6. In the Appendix there are classification result tables for specific test cases illustrating top results by precision.

## 2 Related Work

To our knowledge this was the first time a machine learning approach was attempted to static code analysis with the first results demonstrated during the SATE2010 workshop [Mok10d, Mok12, ODBN10]. In the same year, a somewhat similar approach independently was presented [BSSV10] for vulnerability classification and prediction using machine learning and SVMs, but working with a different set of data.

Additional related work (to various degree of relevance or use) is further listed (this list is not exhaustive). A taxonomy of Linux kernel vulnerability solutions in terms of patches and source code as well as categories for both are found in [MLB07]. The core ideas and principles behind the MARF's pipeline and testing methodology for various algorithms in the pipeline

adapted to this case are found in [Mok08b, Mok10b] as it was the easiest implementation available to accomplish the task. There also one can find the majority of the core options used to set the configuration for the pipeline in terms of algorithms used. A binary analysis using machine learning approach for quick scans for files of known types in a large collection of files is described in [MD08] as well as the NLP and machine learning for NLP tasks in DEFT2010 [Mok10c, Mok10a] with the corresponding DEFT2010App and its predecessor for hand-written image processing WriterIdentApp [MSS09]. Tlili’s 2009 PhD thesis covers topics on automatic detection of safety and security vulnerabilities in open source software [Tli09]. Statistical analysis, ranking, approximation, dealing with uncertainty, and specification inference in static code analysis are found in the works of Engler’s team [KTB<sup>+</sup>06, KAYE04, KE03]. Kong et al. further advance static analysis (using parsing, etc.) and specifications to eliminate human specification from the static code analysis in [KZL10]. Spectral techniques are used for pattern scanning in malware detection by Eto et al. in [ESI<sup>+</sup>09]. Some researchers propose a general data mining system for incident analysis with data mining engines in [IYE<sup>+</sup>09]. Hanna et al. describe a synergy between static and dynamic analysis for the detection of software security vulnerabilities in [HLYD09] paving the way to unify the two analysis methods. Other researchers propose a MEDUSA system for metamorphic malware dynamic analysis using API signatures in [NJG<sup>+</sup>10]. Some of the statistical NLP techniques we used, are described at length in [MS02]. BitBlaze (and its web counterpart, WebBlaze) are other recent types of tools that to static and dynamic binary code analysis for vulnerabilities fast, developed at Berkeley [Son10a, Son10b]. For wavelets, for example, Li et al. [LjXP<sup>+</sup>09] have shown wavelet transforms and  $k$ -means classification can be used to identify communicating applications on a network fast and is relevant to our study of the code in any form, text or binary.

### 3 Data Sets

We use the SAMATE data set to practically validate our approach. The SAMATE reference data set contains C/C++, Java, and PHP language tracks comprising CVE-selected cases as well as stand-alone cases and the large generated synthetic C and Java test cases (CWE-based, with a lot of variants of different known weaknesses). SATE IV expanded some cases from SATE2010 by increasing the version number, and dropped some other cases (e.g. Chrome).

The C/C++ and Java test cases of various client and server OSS software are compilable into the binary and object code, while the synthetic C and Java cases generated for various CWE entries provided for greater scalability testing (also compilable). The CVE-selected cases had a vulnerable version of a software in question with a list of CVEs attached to it, as well as the most known fixed version within the minor revision number. One of the goals for the CVE-based cases is to detect the known weaknesses outlined in CVEs using static code analysis and also to verify if they were really fixed in the “fixed version” [ODBN12]. The cases with known CVEs and CWEs were used as the training models described in the methodology. The summary below is a union of the data sets from SATE2010 and SATE IV.

The preliminary list of the CVEs that the organizers expect to locate in the test cases were collected from the NVD [NIS12a, ODBN12] for Wireshark 1.2.0, Dovecot, Tomcat 5.5.13, Jetty 6.1.16, and Wordpress 2.0.

The specific test cases with versions and language at the time included CVE-selected:

- C: Wireshark 1.2.0 (vulnerable) and Wireshark 1.2.18 (fixed, up from Wireshark 1.2.9 in SATE2010)

- C: Dovecot (vulnerable) and Dovecot (fixed)
- C++: Chrome 5.0.375.54 (vulnerable) and Chrome 5.0.375.70 (fixed)
- Java: Tomcat 5.5.13 (vulnerable) and Tomcat 5.5.33 (fixed, up from Tomcat 5.5.29 in SATE2010)
- Java: Jetty 6.1.16 (vulnerable) and Jetty 6.1.26 (fixed)
- PHP: Wordpress 2.0 (vulnerable) and Wordpress 2.2.3 (fixed)

originally non-CVE selected in SATE2010:

- C: Dovecot
- Java: Pebble 2.5-M2

Synthetic CWE cases produced by the SAMATE team:

- C: Synthetic C covering 118 CWEs and  $\approx 60K$  files
- Java: Synthetic Java covering  $\approx 50$  CWEs and  $\approx 20K$  files

## 4 Methodology

In this section we outline the methodology of our approach to static source code analysis. Most of this methodology is an updated description from [Mok10d]. The line number determination methodology is also detailed in [Mok10d, ODBN10], but is not replicated here. Thus, the methodology’s principles overview is described in Section 4.1, the knowledge base construction is in Section 4.2, machine learning categories in Section 4.3, and the high-level algorithmic description is in Section 4.4.

### 4.1 Methodology Overview

The core methodology principles include:

- Machine learning and dynamic programming
- Spectral and signal processing techniques
- NLP  $n$ -gram and smoothing techniques (add- $\delta$ , Witten-Bell, MLE, etc.)

We use signal processing techniques, i.e. presently we do not parse or otherwise work at the syntax and semantics levels. We treat the source code as a “signal”, equivalent to binary, where each  $n$ -gram ( $n = 2$  presently, i.e. two consecutive characters or, more generally, bytes) are used to construct a sample amplitude value in the signal. In the NLP pipeline, we similarly treat the source code as a “characters”, where each  $n$ -gram ( $n = 1..3$ ) is used to construct the language model.

We show the system the examples of files with weaknesses and MARFCAT learns them by computing spectral signatures using signal processing techniques or various language models (based on options) from CVE-selected test cases. When some of the mentioned techniques

are applied (e.g. filters, silence/noise removal, other preprocessing and feature extraction techniques), the line number information is lost as a part of this process.

When we test, we compute either how similar or distant each file is from the known trained-on weakness-laden files or compare trained language models with the unseen language fragments in the NLP pipeline. In part, the methodology can approximately be seen as some signature-based antivirus or IDS software systems detect bad signature, except that with a large number of machine learning and signal processing algorithms, we test to find out which combination gives the highest precision and best run-time.

At the present, however, we are looking at the whole files instead of parsing the finer-grain details of patches and weak code fragments. This aspect lowers the precision, but is relatively fast to scan all the code files.

## 4.2 CVEs and CWEs – the Knowledge Base

The CVE-selected test cases serve as a source of the knowledge base to gather information of how known weak code “looks like” in the signal form [Mok10d], which we store as spectral signatures clustered per CVE or CWE (Common Weakness Enumeration). The introduction by the SAMATE team of a large synthetic code base with CWEs, serves as a part of knowledge base learning as well. Thus, we:

- Teach the system from the CVE-based cases
- Test on the CVE-based cases
- Test on the non-CVE-based cases

For synthetic cases we do similarly:

- Teach the system from the CWE-based synthetic cases
- Test on the CWE-based synthetic cases
- Test on the CVE and non-CVE-based cases for CWEs from synthetic cases

We create index files in XML in the format similar to that of SATE to index all the file of the test case under study. The CVE-based cases after the initial index generation are manually annotated from the NVD database before being fed to the system. The script that does the initial index gathering in the OSS distribution of MARFCAT is called `collect-files-meta.pl` written in Perl. The synthetic cases required a special modification to that resulting in `collect-files-meta-synthetic.pl` where there are no CVEs to fill in but CWEs alone, with the auto-prefilled explanations since the information in the synthetic cases is not arbitrary and controlled for identification.

## 4.3 Categories for Machine Learning

The tow primary groups of classes we train and test on include are naturally the CVEs [NIS12a, NIS12b] and CWEs [VM12]. The advantages of CVEs is the precision and the associated meta knowledge from [NIS12a, NIS12b] can be all aggregated and used to scan successive versions of the the same software or derived products (e.g. WebKit in multiple browsers). CVEs are also generally uniquely mapped to CWEs. The CWEs as a primary class, however, offer broader

categories, of kinds of weaknesses there may be, but are not yet well assigned and associated with CVEs, so we observe the loss of precision. Since we do not parse, we generally cannot deduce weakness types or even simple-looking aspects like line numbers where the weak code may be. So we resort to the secondary categories, that are usually tied into the first two, which we also machine-learn along, such as issue types (*sink*, *path*, *fix*) and line numbers.

## 4.4 Algorithms

In our methodology we systematically test and select the best (a tradeoff between speed and accuracy) combination(s) of the algorithm implementations available to us and then use only those for subsequent testing. This methodology is augmented with the cases when the knowledge base for the same code type is learned from multiple sources (e.g. several independent C test cases).

### 4.4.1 Signal Pipeline

Algorithmically-speaking, the steps that are performed in the machine-learning signal based analysis are in Figure 1. The specific algorithms come from the classical literature and other sources and are detailed in [Mok08b] and the related works. To be more specific for this work, the loading typically refers to the interpretation of the files being scanned in terms of bytes forming amplitude values in a signal (as an example, 8kHz or 16kHz frequency) using either uni-gram, bi-gram, or tri-gram approach. Then, the preprocessing allows to be none at all (“raw”, or the fastest), normalization, traditional frequency domain filters, wavelet-based filters, etc. Feature extraction involves reducing an arbitrary length signal to a fixed length feature vector of what thought to be the most relevant features are in the signal (e.g. spectral features in FFT, LPC), min-max amplitudes, etc. Classification stage is then separated either to train by learning the incoming feature vectors (usually  $k$ -means clusters, median clusters, or plain feature vector collection, combined with e.g. neural network training) or testing them against the previously learned models.

### 4.4.2 NLP Pipeline

The steps that are performed in NLP and the machine-learning based analysis are presented in Figure 2. The specific algorithms again come from the classical literature (e.g. [MS02]) and are detailed in [Mok10b] and the related works. To be more specific for this work, the loading typically refers to the interpretation of the files being scanned in terms of  $n$ -grams: uni-gram, bi-gram, or tri-gram approach and the associated statistical smoothing algorithms, the results of which (a vector, 2D or 3D matrix) are stored.

## 4.5 Binary and Bytecode Analysis

In this iteration we also perform preliminary Java bytecode and compiled C code static analysis and produce results using the same signal processing, NLP, combined with machine learning and data mining techniques. At this writing, the NIST SAMATE synthetic reference data set for Java and C was used. The algorithms presented in Section 4.4 are used as-is in this scenario with the modifications to the index files. The modifications include removal of the line numbers, source code fragments, and lines-of-text counts (which are largely meaningless and ignored. The byte counts may be recomputed and capturing a byte offset instead of a line number was projected. The filenames of the index files were updated to include `-bin` in them

```

// Construct an index mapping CVEs to files and locations within files
1 Compile meta-XML index files from the CVE reports (line numbers, CVE, CWE,
  fragment size, etc.). Partly done by a Perl script and partly annotated manually;
2 foreach source code base, binary code base do
    // Presently in these experiments we use simple mean clusters of
    // feature vectors or unigram language models per default MARF
    // specification ([Mok08b, The12])
3   Train the system based on the meta index files to build the knowledge base (learn);
4   begin
5     Load (interpret as a wave signal or  $n - gram$ );
6     Preprocess (none, FFT-filters, wavelets, normalization, etc.);
7     Extract features (FFT, LPC, min-max, etc.);
8     Train (Similarity, Distance, Neural Network, etc.);
9   end
10  Test on the training data for the same case (e.g. Tomcat 5.5.13 on Tomcat 5.5.13)
    with the same annotations to make sure the results make sense by being high and
    deduce the best algorithm combinations for the task;
11  begin
12    Load (same);
13    Preprocess (same);
14    Extract features (same);
15    Classify (compare to the trained  $k$ -means, or medians, or language models);
16    Report;
17  end
18  Similarly test on the testing data for the same case (e.g. Tomcat 5.5.13 on Tomcat
    5.5.13) without the annotations as a sanity check;
19  Test on the testing data for the fixed case of the same software (e.g. Tomcat 5.5.13 on
    Tomcat 5.5.33);
20  Test on the testing data for the general non-CVE case (e.g. Tomcat 5.5.13 on Pebble
    or synthetic);
21 end

```

Figure 1: Machine-learning-based static code analysis testing algorithm using the signal pipeline

to differentiate from the original index files describing the source code. Another point is at the moment the simplifying assumption is that each compilable source file e.g. `.java` or `.c` produce the corresponding `.class` and `.o` files that we examine. We do not examine inner classes or linked executables or libraries at this point.

## 4.6 Wavelets

As a part of a collaboration project with Dr. Yankui Sun from Tsinghua University, wavelet-based signal processing for the purposes of noise filtering is being introduced with this work to compare it to no-filtering, or FFT-based classical filtering. It's been also shown in [LjXP<sup>+</sup>09] that wavelet-aided filtering could be used as a fast preprocessing method for a network application identification and traffic analysis [LKW08].



```

1 Compile meta-XML index files from the CVE reports (line numbers, CVE, CWE,
  fragment size, etc.). Partly done by a Perl script and partly annotated manually;
2 foreach source code base, binary code base do
    // Presently in these experiments we use simple unigram language models
    per default MARF specification ([Mok10b])
3   Train the system based on the meta index files to build the knowledge base (learn);
4   begin
5     | Load (n-gram);
6     | Train (statistical smoothing estimators);
7   end
8   Test on the training data for the same case (e.g. Tomcat 5.5.13 on Tomcat 5.5.13)
    with the same annotations to make sure the results make sense by being high and
    deduce the best algorithm combinations for the task;
9   begin
10    | Load (same);
11    | Classify (compare to the trained language models);
12    | Report;
13  end
14  Similarly test on the testing data for the same case (e.g. Tomcat 5.5.13 on Tomcat
    5.5.13) without the annotations as a sanity check;
15  Test on the testing data for the fixed case of the same software (e.g. Tomcat 5.5.13 on
    Tomcat 5.5.33);
16  Test on the testing data for the general non-CVE case (e.g. Tomcat 5.5.13 on Pebble
    or synthetic);
17 end

```

Figure 2: Machine-learning-based static code analysis testing algorithm using the NLP pipeline

We rely in part on the the algorithm and methodology found in [AS01, SCL<sup>+</sup>03, KBC05, KBC06], and at this point only a separating 1D discrete wavelet transform (SDWT) has been tested (see Section 5.4.1).

Since the original wavelet implementation [SCL<sup>+</sup>03] is in MATLAB [Mat12a, Sch07], we used in part the `codegen` tool from the MATLAB Coder toolbox [Mat12b, Mat12c] to generate a rough C/C++ equivalent in order to (manually) translate some fragments into Java (the language of MARF and MARFCAT). The specific function for up/down sampling used by the wavelets function in [Mot09] written also C/C++ was translated to Java in MARF as well with unit tests added.

#### 4.7 Demand-Driven Distributed Evaluation with GIPSY

To enhance the scalability of the approach, we convert the MARFCAT stand-alone application to a distributed one using an educative model of computation (demand-driven) implemented in the General Intensional Programming System (GIPSY)’s multi-tier run-time system [Han10, Ji11, Vas05, Paq09], which can be executed distributively using Jini (Apache River), or JMS [JMP12].

To adapt the application to the GIPSY’s multi-tier architecture, we create a problem-specific generator and worker tiers (PS-DGT and PS-DWT respectively) for the MARFCAT application.

The generator(s) produce demands of what needs to be computed in the form of a file (source code file or a compiled binary) to be evaluated and deposit such demands into a store managed by the demand store tier (DST) as pending. Workers pickup pending demands from the store, and then process them (all tiers run on multiple nodes) using a traditional MARFCAT instance. Once the result (a **Warning** instance) is computed, the PS-DWT deposit it back into the store with the status set to *computed*. The generator “harvests” all computed results (warnings) and produces the final report for a test cases. Multiple test cases can be evaluated simultaneously or a single case can be evaluated distributively. This approach helps to cope with large amounts of data and avoid recomputing warnings that have already been computed and cached in the DST.

The initial basic experiment assumes the PS-DWTs have the training sets data and the test cases available to them from the start (either by a copy or via an NFS/CIFS-mounted volumes); thus, the distributed evaluation only concerns with the classification task only as of this version. The follow up work will remove this limitation.

In this setup a demand represents a file (a path) to scan (actually a an instance of the `FileItem` object), which is deposited into the DST. The PS-DWT picks up that and checks the file per training set that’s already there and returns a **ResultSet** object back into the DST under the same demand signature that was used to deposit the path to scan. The result set is sorted from the most likely to the list likely with a value corresponding to the distance or similarity. The PS-DGT picks up the result sets and does the final output aggregation and saves report in one of the desired report formats (see Section 4.8 picking up the top two results from the result set and testing against a threshold to accept or reject the file (path) as vulnerable or not. This effectively splits the monolithic MARFCAT application in two halves in distributing the work to do where the classification half is arbitrary parallel.

Simplifying assumptions:

- Test case data and training sets are present on each node (physical or virtual) in advance (via a copy or a CIFS or NFS volume), so no demand driven training occurs, only classification
- The demand assumes to contain only file information to be examined (`FileItem`)
- PS-DWT assumes a single pre-defined configuration, i.e. configuration for MARFCAT’s option is not a part of the demand
- PS-DWT assume CVE or CWE testing based on its local settings and not via the configuration in a demand

## 4.8 Export

### 4.8.1 SATE

By default MARFCAT produces the report data in the SATE XML format, according to the SATE IV requirements. In this iteration other formats are being considered and realized. To enable multiple format output, the MARFCAT report generation data structures were adapted case-based output.

### 4.8.2 Forensic Lucid

The first one, is Forensic Lucid, the author Mokhov’s PhD topic, a language to specify and evaluate digital forensic cases by uniformly encoding the evidence and witness accounts (eviden-

tial statement or knowledge base) of any case from multiple sources (system specs, logs, human accounts, etc.) as a description of an incident to further perform investigation and event reconstruction. Following the data export in Forensic Lucid in the preceding work [MPD08, MPD10, Mok08a] we use it as a format for evidential processing of the results produced by MARFCAT. The work [MPD08] provides details of the language; it will suffice to mention here that the report generated by MARFCAT in Forensic Lucid is a collection of warnings as observations with the hierarchical notion of nested context of warning and location information. These will form an evidential statement in Forensic Lucid. The example scenario where such evidence compiled via a MARFCAT Forensic Lucid report would be in web-based applications and web browser-based incident investigations of fraud, XSS, buffer overflows, etc. linking CVE/CWE-based evidence analysis of the code (binary or source) security bugs with the associated web-based malware propagation or attacks to provide possible events where specific attacks can be traced back to the specific security vulnerabilities.

### 4.8.3 SAFES

The third format, for which the export functionality is not done as of this writing, SAFES, is the 3rd format for output of the MARFCAT. SAFES is becoming a standard to reporting such information and the SATE organizers began endorsing it as an alternative during SATE IV.

## 4.9 Experiments

The below is the current summary of the conducted experiments:

- Re-testing of the newer fixed versions such as Wireshark 1.2.18 and Tomcat 5.5.33.
- Half-based testing of the previous versions by reducing the training set by half and but testing for all known CVEs or CWEs for Wireshark 1.2.18, Tomcat 5.5.33, and Chrome 5.0.375.54.
- Testing the new test cases of Dovecot, Jetty 6.1.x, and Wordpress 2.x as well as Synthetic C and Synthetic Java.
- Binary test on the Synthetic C and Synthetic Java test cases.
- Performing tests using wavelets for preprocessing.

## 5 Results

The preliminary results of application of our methodology are outlined in this section. We summarize the top precisions per test case using either signal-processing or NLP-processing of the CVE-based and synthetic cases and their application to the general cases. Subsequent sections detail some of the findings and issues of MARFCAT's result releases with different versions. Some experiments we compare the results with the previously obtained ones [Mok10d] where compatible and appropriate.

The results currently are being gradually released in the iterative manner that were obtained through the corresponding versions of MARFCAT as it was being designed and developed.

## 5.1 Preliminary Results Summary

The results summarize the half-training-full-testing data vs. that of regular ones reported in [Mok10d].

- Wireshark:
  - CVEs (signal): 92.68%, CWEs (signal): 86.11%,
  - CVEs (NLP): 83.33%, CWEs (NLP): 58.33%
- Tomcat:
  - CVEs (signal): 83.72%, CWEs (signal): 81.82%,
  - CVEs (NLP): 87.88%, CWEs (NLP): 39.39%
- Chrome:
  - CVEs (signal): 90.91%, CWEs (signal): 100.00%,
  - CVEs (NLP): 100.00%, CWEs (NLP): 88.89%
- Dovecot:
  - 14 warnings; but it appears all quality or false positive
  - (very hard to follow the code, severely undocumented)
- Pebble:
  - none found during quick testing
- Wireshark:
  - CVEs (signal): 92.68%, CWEs (signal): 86.11%,
  - CVEs (NLP): 83.33%, CWEs (NLP): 58.33%
- Tomcat:
  - CVEs (signal): 83.72%, CWEs (signal): 81.82%,
  - CVEs (NLP): 87.88%, CWEs (NLP): 39.39%
- Dovecot 1.2.x: (ongoing of this writing)
- Jetty: (ongoing of this writing)
- Wordpress: (ongoing of this writing)
- Chrome:
  - CVEs (signal): 90.91%, CWEs (signal): 100.00%,
  - CVEs (NLP): 100.00%, CWEs (NLP): 88.89%
- Dovecot (new, 2.x):
  - 14 warnings; but it appears all quality or false positive

- (very hard to follow the code, severely undocumented)
- Pebble:
  - none found during quick testing

What follows are some select statistical measurements of the precision in recognizing CVEs and CWEs under different configurations using the signal processing and NLP processing techniques.

“Second guess” statistics provided to see if the hypothesis that if our first estimate of a CVE/CWE is incorrect, the next one in line is probably the correct one. Both are counted if the first guess is correct.

A sample signal visuiasalization in the middle of a vulnerable file `packet-afs.c` in Wireshark 1.2.0 to CVE-2009-2562 is in Figure 3 in the wave form. The low “dips” represent the text line endings (coupled with a preceding character (bytes) in bigrams (two PCM-signed bytes assumed encoded in 8kHz representing the amplitude; normalized), which are often either semicolons, closing or opening braces, brackets or parentheses). Only a small fragment is shown of roughly 300 bytes in length to be visually comprehensive of a nature of a signal we a dealing with.

In Figure 4, there are 3 spectrograms generated for the same file `packet-afs.c`. The first two columns represent the CVE-2009-2562-vulnerable file, both versions are the same with ehanced contrast to see the detail. The subsequent pairs are of the same file in Wireshark 1.2.9 and Wireshark 1.2.18, where CVE-2009-2562 is no longer present. Small changes are noticeable primarily in the bottom left and top right corners of the images, and even smaller elsewhere in the images.

## 5.2 Version SATE-IV.1

### 5.2.1 Half-Training Data For Training and Full For Testing

This is one of the experiment per discussion with Aurelien Delaitre and SATE organizers. The main idea is to test robustness and precision of the MARFCAT approach by artificially reducing known weaknesses (their locations) to learn from by 50%, but test on the whole 100% to see how much does precision degrade with such a reduction.

Supplying only CWE classes testing for this experiment (CVE classes make little sense). Only the first 50% of the entries entries were used for training for Wireshark 1.2.0, Tomcat 5.5.13, and Chrome 5.0.375.54, while the full 100% were used to test the precision changes. The below are the results.

It should be noted that CWE classification is generally less accurate due to lots of things stuffed (by NVD) into very broad categories such as NVD-CWE-Other and NVD-CWE-noinfo. Additionally, since we arbitrarily picked the first 50% of the training data, some of the CWEs simply were left out completely and not trained on if they were entirely in the omitted half, so their individual precision is obviously 0% when tested for.

The archive contains the .log and the .xml files (the latter for now are in SATE format only with the scientific notation +E3 removed). The best reports are:

```
report-cweidnoprepreprawfftcheb-wireshark-1.2.0-half-train-cwe.xml
report-cweidnoprepreprawfftdiff-wireshark-1.2.0-half-train-cwe.xml
```

The experiments are subdivided into regular (signal) and NLP based testing.

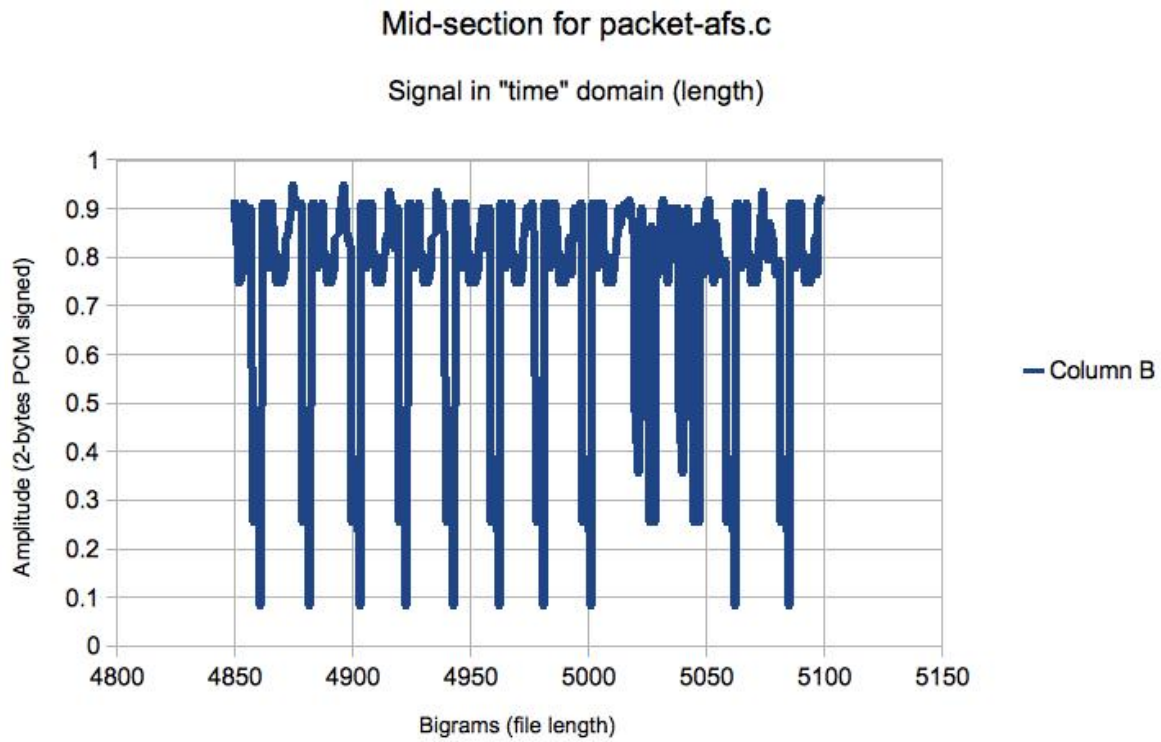


Figure 3: A wave graph of a fraction of the CVE-2009-2562-vulnerable `packet-afs.c` in Wireshark 1.2.0

### Signal.

- Wireshark 1.2.0:

Reduction of the training data by half resulted in  $\approx 14\%$  precision drop compared to the previous result (best 86.11% see the NIST report [Mok11], vs. 72.22% overall).

New results (by algorithms, then by CWEs):

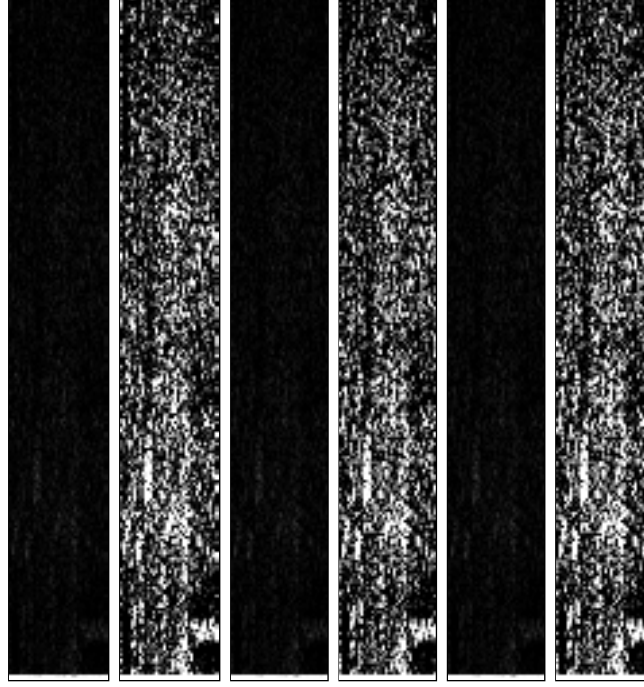


Figure 4: Spectrograms of CVE-2009-2562-vulnerable `packet-afs.c` in Wireshark 1.2.0, fixed Wireshark 1.2.9 and Wireshark 1.2.18

guess	run	algorithms	good	bad	%
1st	1	-cweid -nopreprep -raw -fft -cheb	26	10	72.22
1st	2	-cweid -nopreprep -raw -fft -diff	26	10	72.22
1st	3	-cweid -nopreprep -raw -fft -eucl	22	14	61.11
1st	4	-cweid -nopreprep -raw -fft -cos	25	23	52.08
1st	5	-cweid -nopreprep -raw -fft -mink	17	19	47.22
1st	6	-cweid -nopreprep -raw -fft -hamming	17	19	47.22
2nd	1	-cweid -nopreprep -raw -fft -cheb	30	6	83.33
2nd	2	-cweid -nopreprep -raw -fft -diff	30	6	83.33
2nd	3	-cweid -nopreprep -raw -fft -eucl	24	12	66.67
2nd	4	-cweid -nopreprep -raw -fft -cos	32	16	66.67
2nd	5	-cweid -nopreprep -raw -fft -mink	23	13	63.89
2nd	6	-cweid -nopreprep -raw -fft -hamming	24	12	66.67
guess	run	class	good	bad	%
1st	1	NVD-CWE-noinfo	68	39	63.55
1st	2	CWE-20	38	22	63.33
1st	3	CWE-119	18	14	56.25
1st	4	NVD-CWE-Other	9	8	52.94
1st	5	CWE-189	0	12	0.00
2nd	1	NVD-CWE-noinfo	84	23	78.50
2nd	2	CWE-20	39	21	65.00
2nd	3	CWE-119	29	3	90.62
2nd	4	NVD-CWE-Other	11	6	64.71
2nd	5	CWE-189	0	12	0.00

- Tomcat 5.5.13:

Drop from 81.82% (see NIST report’s Table 7, p. 70) to 75% top result as a result (about 7 points) of training data reduction by 50%.

New precision estimates:

guess	run	algorithms	good	bad	%
1st	1	-cweid -nopreprep -raw -fft -diff	6	2	75.00
1st	2	-cweid -nopreprep -raw -fft -hamming	5	9	35.71
2nd	1	-cweid -nopreprep -raw -fft -diff	6	2	75.00
2nd	2	-cweid -nopreprep -raw -fft -hamming	8	6	57.14
guess	run	class	good	bad	%
1st	1	CWE-264	1	0	100.00
1st	2	CWE-255	2	0	100.00
1st	3	CWE-200	1	0	100.00
1st	4	CWE-22	6	3	66.67
1st	5	CWE-79	1	4	20.00
1st	6	CWE-119	0	2	0.00
1st	7	CWE-20	0	2	0.00
2nd	1	CWE-264	1	0	100.00
2nd	2	CWE-255	2	0	100.00
2nd	3	CWE-200	1	0	100.00
2nd	4	CWE-22	7	2	77.78
2nd	5	CWE-79	3	2	60.00
2nd	6	CWE-119	0	2	0.00
2nd	7	CWE-20	0	2	0.00

- Chrome 5.0.375.54:

Chrome result is for completeness even though it is not a test case for SATE IV.

Chrome is poor for some reason – drop from 100% (Table 5, p. 68) to 44.44%, but it’s only 9 entries. The first result below is invalid, i.e. with a poor recall (the sum of  $2 + 0 < 9$ , should be total 9; I haven’t looked at yet as of why).



guess	run	algorithms	good	bad	%
1st	1	-cweid -nopreprep -raw -fft -cos	2	0	100.00
1st	2	-cweid -nopreprep -raw -fft -eucl	4	5	44.44
1st	3	-cweid -nopreprep -raw -fft -cheb	3	6	33.33
1st	4	-cweid -nopreprep -raw -fft -hamming	3	6	33.33
1st	5	-cweid -nopreprep -raw -fft -mink	2	7	22.22
2nd	1	-cweid -nopreprep -raw -fft -cos	2	0	100.00
2nd	2	-cweid -nopreprep -raw -fft -eucl	4	5	44.44
2nd	3	-cweid -nopreprep -raw -fft -cheb	4	5	44.44
2nd	4	-cweid -nopreprep -raw -fft -hamming	4	5	44.44
2nd	5	-cweid -nopreprep -raw -fft -mink	3	6	33.33
guess	run	class	good	bad	%
1st	1	CWE-94	6	3	66.67
1st	2	CWE-20	3	2	60.00
1st	3	CWE-79	2	2	50.00
1st	4	NVD-CWE-noinfo	2	2	50.00
1st	5	NVD-CWE-Other	1	7	12.50
1st	6	CWE-399	0	4	0.00
1st	7	CWE-119	0	4	0.00
2nd	1	CWE-94	6	3	66.67
2nd	2	CWE-20	3	2	60.00
2nd	3	CWE-79	3	1	75.00
2nd	4	NVD-CWE-noinfo	3	1	75.00
2nd	5	NVD-CWE-Other	2	6	25.00
2nd	6	CWE-399	0	4	0.00
2nd	7	CWE-119	0	4	0.00

**NLP.** Generally this genre of classification was poor as before in this experiment, all around 40-45% percent precision, but:

- Wireshark 1.2.0:

New results (by algos, then by CWEs):

guess	run	algorithms	good	bad	%
1st	1	-cweid -nopreprep -char -unigram -add-delta	15	21	41.67
2nd	1	-cweid -nopreprep -char -unigram -add-delta	23	13	63.89
guess	run	class	good	bad	%
1st	1	NVD-CWE-noinfo	11	7	61.11
1st	2	NVD-CWE-Other	1	1	50.00
1st	3	CWE-119	2	3	40.00
1st	4	CWE-20	1	9	10.00
1st	5	CWE-189	0	1	0.00
2nd	1	NVD-CWE-noinfo	17	1	94.44
2nd	2	NVD-CWE-Other	1	1	50.00
2nd	3	CWE-119	4	1	80.00
2nd	4	CWE-20	1	9	10.00
2nd	5	CWE-189	0	1	0.00

- Tomcat 5.5.13:

Strangely, the best result is higher than with all of the data in the past report (42.42% below vs. previous 39.39%).

guess	run	algorithms	good	bad	%
1st	1	-cweid -nopreprep -char -unigram -add-delta	14	19	42.42
2nd	1	-cweid -nopreprep -char -unigram -add-delta	18	15	54.55
guess	run	class	good	bad	%
1st	1	CWE-255	1	0	100.00
1st	2	CWE-264	2	0	100.00
1st	3	CWE-119	1	0	100.00
1st	4	CWE-20	1	0	100.00
1st	5	CWE-22	7	9	43.75
1st	6	CWE-200	1	3	25.00
1st	7	CWE-79	1	6	14.29
1st	8	CWE-16	0	1	0.00
2nd	1	CWE-255	1	0	100.00
2nd	2	CWE-264	2	0	100.00
2nd	3	CWE-119	1	0	100.00
2nd	4	CWE-20	1	0	100.00
2nd	5	CWE-22	11	5	68.75
2nd	6	CWE-200	1	3	25.00
2nd	7	CWE-79	1	6	14.29
2nd	8	CWE-16	0	1	0.00

- Chrome 5.0.375.54:

Here drop is twice as much ( $\approx 44\%$  vs.  $88\%$ ).

guess	run	algorithms	good	bad	%
1st	1	-cweid -nopreprep -char -unigram -add-delta	4	5	44.44
2nd	1	-cweid -nopreprep -char -unigram -add-delta	5	4	55.56
guess	run	class	good	bad	%
1st	1	NVD-CWE-noinfo	1	0	100.00
1st	2	CWE-79	1	0	100.00
1st	3	CWE-20	1	0	100.00
1st	4	CWE-94	1	1	50.00
1st	5	CWE-399	0	1	0.00
1st	6	NVD-CWE-Other	0	2	0.00
1st	7	CWE-119	0	1	0.00
2nd	1	NVD-CWE-noinfo	1	0	100.00
2nd	2	CWE-79	1	0	100.00
2nd	3	CWE-20	1	0	100.00
2nd	4	CWE-94	1	1	50.00
2nd	5	CWE-399	0	1	0.00
2nd	6	NVD-CWE-Other	0	2	0.00
2nd	7	CWE-119	1	0	100.00

### 5.3 Version SATE-IV.2

These runs represent using the same SATE2010 training data for Tomcat 5.5.13 Wireshark 1.2.0 to test the updated fixed versions (as from SATE2010) to Tomcat 5.5.33 and Wireshark 1.2.18 using the same settings. At this run, no new CVEs that may have happened from the previous fixed versions of Tomcat 5.5.29 and Wireshark 1.2.9 respectively in 2010 were added to the training data for the versions being tested in this experiment as to see if any old issues reoccur or not. In this short summary, both signal and NLP testing reveal no same known issues found.

- SATE-IV.2-train-test-test-run-quick-tomcat-5-5-33-cve

This is CVE-based classical signal classification.

A typical MARFCAT run. Tomcat 5.5.13 used for training. For most reports, no warnings were spotted based on what was learned from 5.5.13, so the reports convey earlier CVEs were fixed.

Empty reports like:

```
report-noprepreprawfftcheb-train-test-test-run-quick-tomcat-5-5-33-cve.xml
```

However, the `-cos` report is noisy and non-empty:

```
report-noprepreprawfftcos-train-test-test-run-quick-tomcat-5-5-33-cve.xml
```

Overly detailed log files are also provided.

- SATE-IV.2-train-test-test-run-quick-tomcat-5-5-33-cwe

This is classical CWE-based testing.

A typical MARFCAT CWE run. Tomcat 5.5.13 used for training.

No warnings found based on the CVE data learned.

Most of the reports are empty, e.g.:

```
report-nopreprepcharunigramadddelta-train-test-test-run-quick-tomcat-5-5-33-cve-nlp.xml
```

The `-cos` report is not as noisy as for CVEs, but still contains a couple of false positives.

```
report-cweidnoprepreprawfftcos-train-test-test-run-quick-tomcat-5-5-33-cwe.xml
```

Overly detailed log files also provided.

Training and testing indexes are provided (`*_test.xml` and `*_train.xml`).

- SATE-IV.2-train-test-test-run-quick-tomcat-5-5-33-cve-nlp

This is CVE-based NLP testing.

A typical MARFCAT NLP run. Tomcat 5.5.13 used for training. Usually a slow run, so only one configuration is tried. No warnings found based on the CVE data learned.

The only empty report is:

```
report-nopreprepcharunigramadddelta-train-test-test-run-quick-tomcat-5-5-33-cve-nlp.xml
```

However, the `-cos` report is noisy and non-empty:

```
report-noprepreprawfftcos-train-test-test-run-quick-tomcat-5-5-33-cve.xml
```

Overly detailed log files also provided.

Training and testing indexes are provided (\*\_test.xml and \*\_train.xml).

- SATE-IV.2-train-test-test-run-quick-tomcat-5-5-33-cwe-nlp

This is CWE-based NLP testing.

A typical MARFCAT CWE NLP run. Tomcat 5.5.13 used for training. Usually a slow run, so only one configuration is tried. No warnings found based on the CVE data learned.

The only empty report is:

```
report-cweidnopreprepcharunigramadddelta-train-test-test-run-quick-tomcat-5-5-33-cwe-nlp.xml
```

Overly detailed log files also provided.

- SATE-IV.2-train-test-test-run-quick-wireshark-1-2-18-cve

Test Wireshark 1.2.18 using the training data from Wireshark 1.2.0 and classical CVE-based processing.

Majority of algorithms returned empty reports. -cos was as noisy as usual, but -mink was non-empty but quite short (though also presumed with false positives).

Empty reports:

```
report-noprepreprawfftcheb-train-test-test-run-quick-wireshark-1-2-18-cve.xml
report-noprepreprawfftdiff-train-test-test-run-quick-wireshark-1-2-18-cve.xml
report-noprepreprawffteucl-train-test-test-run-quick-wireshark-1-2-18-cve.xml
report-noprepreprawffthamming-train-test-test-run-quick-wireshark-1-2-18-cve.xml
```

Non empty reports:

```
report-noprepreprawfftcos-train-test-test-run-quick-wireshark-1-2-18-cve.xml
report-noprepreprawfftmink-train-test-test-run-quick-wireshark-1-2-18-cve.xml
```

Verbose log files and input index files are also supplied for the most cases.

[TODO]

## 5.4 Version SATE-IV.5

### 5.4.1 Wavelet Experiments

The preliminary experiments using the separating discreet wavelet transform (DWT) filter are summarized in Table 1 and Table 3 for CVEs and CWEs respectively. For comparison, the low-pass FFT filter is used for the same as shown in Table 2 and Table 4 respectively. For the CVE experiments, the wavelet transforms overall produces better precision across configurations (larger number of configurations produce higher precision result) than those with the low-pass FFT filter. While the top precision result remains the same, it is shown than when filtering is wanted, the wavelet transform is perhaps a better choice for some configurations, e.g. from 4 and below as well as for the 2nd guess statistics. The very top result for the CWE based processing so far exceeds the overall precision of separating DWT vs. low-pass FFT, which then drops below for the subsequent configurations. -cos was dropped from Table 3 for technical reasons. In Figure 5 is a spectrogram with the SDWT preprocessing in the pipeline. More exploration

in this area is under way for more advanced wavelet filters than the simple separating DWT filter as to see whether they would outperform `-raw` or not and at the same time minimizing the run-time performance decrease with the extra filtering.



Figure 5: A spectrogram of CVE-2009-2562-vulnerable `packet-afs.c` in Wireshark 1.2.0, after SDWT

## 6 Conclusion

We review the current results of this experimental work, its current shortcomings, advantages, and practical implications.

### 6.1 Shortcomings

The below is a list of most prominent issues with the presented approach. Some of them are more “permanent”, while others are solvable and intended to be addressed in the future work. Specifically:

- Looking at a signal is less intuitive visually for code analysis by humans. (However, can produce a problematic “spectrogram” in some cases).
- Line numbers are a problem (easily “filtered out” as high-frequency “noise”, etc.). A whole “relativistic” and machine learning methodology developed for the line numbers in [Mok10d] to compensate for that. Generally, when CVEs are the primary class, by accurately identifying the CVE number one can get all the other pertinent details from the CVE database, including patches and line numbers making this a lesser issue.

- Accuracy depends on the quality of the knowledge base (see Section 4.2) collected. Some of this collection and annotation is manual to get the indexes right, and, hence, error prone. “Garbage in – garbage out.”
- To detect more of the useful CVE or CWE signatures in non-CVE and non-CWE cases requires large knowledge bases (human-intensive to collect), which can perhaps be shared by different vendors via a common format, such as SATE, SAFES or Forensic Lucid.
- No path tracing (since no parsing is present); no slicing, semantic annotations, context, locality of reference, etc. The “sink”, “path”, and “fix” results in the reports also have to be machine-learned.
- A lot of algorithms and their combinations to try (currently  $\approx 1800$  permutations) to get the best top N. This is, however, also an advantage of the approach as the underlying framework can quickly allow for such testing.
- File-level training vs. fragment-level training – presently the classes are trained based on the entire file where weaknesses are found instead of the known file fragments from CVE-reported patches. The latter would be more fine-grained and precise than whole-file classification, but slower. However, overall the file-level processing is a man-hour limitation than a technological one.
- Separating wavelet filter performance is rather adversely affects the precision to low levels.
- No nice GUI. Presently the application is script/command-line based.

## 6.2 Advantages

There are some key advantages of the approach presented. Some of them follow:

- Relatively fast (e.g. Wireshark’s  $\approx 2400$  files train and test in about 3 minutes) on a now-commodity desktop or a laptop.
- Language-independent (no parsing) – given enough examples can apply to any language, i.e. methodology is the same no matter C, C++, Java or any other source or binary languages (PHP, C#, VB, Perl, bytecode, assembly, etc.) are used.
- Can automatically learn a large knowledge base to test on known and unknown cases.
- Can be used to quickly pre-scan projects for further analysis by humans or other tools that do in-depth semantic analysis as a means to prioritize.
- Can learn from SATE’08, SATE’09, SATE’10, and SATE IV reports.
- Generally, high precision (and recall) in CVE and CWE detection, even at the file level.
- A lot of algorithms and their combinations to select the best for a particular task or class (see Section 4.3).
- Can cope with altered code or code used in other projects (e.g. a lot of problems in Chrome were found in WebKit, used by several browsers).

### 6.3 Practical Implications

Most practical implications of all static code analyzers are obvious – to detect and report source code weaknesses and report them appropriately to the developers. We outline additional implications this approach brings to the arsenal below:

- The approach can be used on any target language without modifications to the methodology or knowing the syntax of the language. Thus, it scales to any popular and new language analysis with a very small amount of effort.
- The approach can nearly identically be transposed onto the compiled binaries and byte-code, detecting vulnerable deployments and installations – sort of like virus scanning of binaries, but instead scanning for infected binaries, one would scan for security-weak binaries on site deployments to alert system administrators to upgrade their packages. XXX: The experiments in this area are ongoing.
- Can learn from binary signatures from other tools like Snort [Sou12].
- The approach is easily extendable to the embedded code and mission-critical code found in aircraft, spacecraft, and various autonomous systems.

### 6.4 Future Work

There is a great number of possibilities in the future work. This includes improvements to the code base of MARFCAT as well as resolving unfinished scenarios and results, addressing shortcomings in Section 6.1, testing more algorithms and combinations from the related work, and moving onto other programming languages (e.g. ASP, C#). Furthermore, plan to conceive collaboration with vendors such as VeraCode, Coverity, and others who have vast data sets to test the full potential of the approach with the others and a community as a whole. Then move on to dynamic code analysis as well applying similar techniques there.

There is a great number of possibilities in the future work. This includes resolving unfinished scenarios and results, addressing shortcomings in Section 6.1, testing more algorithms and combinations from the related work, and moving onto other programming languages (e.g. ASP, C#). Furthermore, foster collaboration with the academic, industry and government vendors that may have vast data sets to test the full potential of the approach with the others and a community as a whole. Then, move on to dynamic code analysis as well applying similar techniques there. Other near-future work items include realization of the SVM-based classification, data export in SAFES and Forensic Lucid formats, a lot of wavelet filtering improvements, and distributed GIPSY cluster-based evaluation.

To improve detection and classification of the malware in the network traffic or otherwise we employ machine learning approach to static pcap payload malicious code analysis and fingerprinting using the open-source MARF framework and its MARFCAT application, originally designed for the SATE static analysis tool exposition workshop. We first train on the known malware pcap data and measure the precision and then test it on the unseen, but known data and select the best available machine learning combination to do so. This work elaborates on the details of the methodology and the corresponding results of application of the machine learning techniques along with signal processing and NLP alike to static network packet analysis in search for malicious code in the packet capture (pcap) data. malicious code analysis [BOB<sup>+</sup>10, SEZS01, SXC<sup>+</sup>04, HJ07, HRSS07, Sue07, RM08, BOA<sup>+</sup>07] We show the system the

examples of pcap files with malware and MARFCAT learns them by computing spectral signatures using signal processing techniques. When we test, we compute how similar or distant each file is from the known trained-on malware-laden files. In part, the methodology can approximately be seen as some signature-based “antivirus” or IDS software systems detect bad signature, except that with a large number of machine learning and signal processing algorithms, we test to find out which combination gives the highest precision and best run-time. At the present, however, we are looking at the whole pcap files. This aspect lowers the precision, but is fast to scan all the files. The malware database with known malware, the reports, etc. serves as a knowledge base to machine-learn from. Thus, we primarily:

- Teach the system from the known cases of malware from their pcap data
- Test on the known cases
- Test on the unseen cases

## 6.5 Acknowledgments

The authors would like to express thanks and gratitude to the following for their help, resources, advice, and otherwise support and assistance:

- NIST SAMATE group
- Dr. Brigitte Jaumard
- Sleiman Rabah
- Open-Source Community

This work is partially supported by the Faculty of ENCS, Concordia University, NSERC, and the 2011-2012 CCSEP scholarship. The wavelet-related work of Yankui Sun is partially supported by the National Natural Science Foundation of China (No. 60971006).

## References

- [AS01] A. F. Abdelnour and I. W. Selesnick. Nearly symmetric orthogonal wavelet bases. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, May 2001.
- [BOA<sup>+</sup>07] M. Bailey, J. Oberheide, J. Andersen, Z. M. Mao, F. Jahanian, and J. Nazario. Automated classification and analysis of Internet malware. Technical report, University of Michigan, April 2007. <http://www.eecs.umich.edu/techreports/cse/2007/CSE-TR-530-07.pdf>.
- [BOB<sup>+</sup>10] Hamad Binsalleeh, Thomas Ormerod, Amine Boukhtouta, Prosenjit Sinha, Amr M. Youssef, Mourad Debbabi, and Lingyu Wang. On the analysis of the zeus botnet crimeware toolkit. In *Eighth Annual Conference on Privacy, Security and Trust, PST 2010, August 17-19, 2010, Ottawa, Ontario, Canada*, pages 31–38. IEEE, 2010.
- [BSSV10] Mehran Bozorgi, Lawrence K. Saul, Stefan Savage, and Geoffrey M. Voelker. Beyond heuristics: Learning to classify vulnerabilities and predict exploits. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge Discovery and Data Mining, KDD’10*, pages 105–114, New York, NY, USA, 2010. ACM.
- [ESI<sup>+</sup>09] Masashi Eto, Kotaro Sonoda, Daisuke Inoue, Katsunari Yoshioka, and Koji Nakao. A proposal of malware distinction method based on scan patterns using spectrum analysis. In *Proceedings of the 16th International Conference on Neural Information Processing: Part II, ICONIP’09*, pages 565–572, Berlin, Heidelberg, 2009. Springer-Verlag.



- [Han10] Bin Han. Towards a multi-tier runtime system for GIPSY. Master's thesis, Department of Computer Science and Software Engineering, Concordia University, Montreal, Canada, 2010.
- [HJ07] K. Hwang and D. Jung. Anti-malware expert system. In H. Martin, editor, *Proceedings of the 17th Virus Bulletin International Conference*, pages 9–17, Vienna, Austria: The Pentagon, Abingdon, OX143YP, England, September 2007.
- [HLYD09] Aiman Hanna, Hai Zhou Ling, Xiaochun Yang, and Mourad Debbabi. A synergy between static and dynamic analysis for the detection of software security vulnerabilities. In Robert Meersman, Tharam S. Dillon, and Pilar Herrero, editors, *OTM Conferences (2)*, volume 5871 of *Lecture Notes in Computer Science*, pages 815–832. Springer, 2009.
- [HRSS07] N. Hnatiw, T. Robinson, C. Sheehan, and N. Suan. Pimp my PE: Parsing malicious and malformed executables. In H. Martin, editor, *Proceedings of the 17th Virus Bulletin International Conference*, pages 9–17, Vienna, Austria: The Pentagon, Abingdon, OX143YP, England, September 2007.
- [IYE<sup>+</sup>09] Daisuke Inoue, Katsunari Yoshioka, Masashi Eto, Masaya Yamagata, Eisuke Nishino, Jun'ichi Takeuchi, Kazuya Ohkouchi, and Koji Nakao. An incident analysis system NICTER and its analysis engines based on data mining techniques. In *Proceedings of the 15th International Conference on Advances in Neuro-Information Processing – Volume Part I*, ICONIP'08, pages 579–586, Berlin, Heidelberg, 2009. Springer-Verlag.
- [Ji11] Yi Ji. Scalability evaluation of the GIPSY runtime system. Master's thesis, Department of Computer Science and Software Engineering, Concordia University, Montreal, Canada, March 2011.
- [JMP12] Yi Ji, Serguei A. Mokhov, and Joey Paquet. Unifying and refactoring DMF to support concurrent Jini and JMS DMS in GIPSY. In Bipin C. Desai, Sudhir P. Mudur, and Emil I. Vassev, editors, *Proceedings of C3S2E 2012*, pages 36–44. ACM, June 2010–2012. Accepted for publication at C3S2E 2012 (to appear); online pre-print <http://arxiv.org/abs/1012.2860>.
- [KAYE04] Ted Kremenek, Ken Ashcraft, Junfeng Yang, and Dawson Engler. Correlation exploitation in error ranking. In *Foundations of Software Engineering (FSE)*, 2004.
- [KBC05] Manesh Kokare, P. K. Biswas, and B. N. Chatterji. Texture image retrieval using new rotated complex wavelet filters. *IEEE Transaction on Systems, Man, and Cybernetics-Part B: Cybernetics*, 6(35):1168–1178, 2005.
- [KBC06] Manesh Kokare, P. K. Biswas, and B. N. Chatterji. Rotation-invariant texture image retrieval using rotated complex wavelet filters. *IEEE Transaction on Systems, Man, and Cybernetics-Part B: Cybernetics*, 6(36):1273–1282, 2006.
- [KE03] Ted Kremenek and Dawson Engler. Z-ranking: Using statistical analysis to counter the impact of static analysis approximations. In *SAS 2003*, 2003.
- [KTB<sup>+</sup>06] Ted Kremenek, Paul Twohey, Godmar Back, Andrew Ng, and Dawson Engler. From uncertainty to belief: Inferring the specification within. In *Proceedings of the 7th Symposium on Operating System Design and Implementation*, 2006.
- [KZL10] Ying Kong, Yuqing Zhang, and Qixu Liu. Eliminating human specification in static analysis. In *Proceedings of the 13th international conference on Recent advances in intrusion detection*, RAID'10, pages 494–495, Berlin, Heidelberg, 2010. Springer-Verlag.
- [LjXP<sup>+</sup>09] Ru Li, Ou jie Xi, Bin Pang, Jiao Shen, and Chun-Lei Ren. Network application identification based on wavelet transform and k-means algorithm. In *Proceedings of the IEEE International Conference on Intelligent Computing and Intelligent Systems (ICIS2009)*, volume 1, pages 38–41, November 2009.
- [LKW08] Kriangkrai Limthong, Fukuda Kensuke, and Pirawat Watanapongse. Wavelet-based unwanted traffic time series analysis. In *2008 International Conference on Computer and Electrical Engineering*, pages 445–449. IEEE Computer Society, 2008.
- [Mat12a] MathWorks. MATLAB. [online], 2000–2012. <http://www.mathworks.com/products/matlab/>.
- [Mat12b] MathWorks. MATLAB Coder. [online], 2012. <http://www.mathworks.com/help/toolbox/>

- `coder/coder_product_page.html`, last viewed June 2012.
- [Mat12c] MathWorks. MATLAB Coder: `codegen` – generate C/C++ code from MATLAB code. [online], 2012. <http://www.mathworks.com/help/toolbox/coder/ref/codegen.html>, last viewed June 2012.
  - [MD08] Serguei A. Mokhov and Mourad Debbabi. File type analysis using signal processing techniques and machine learning vs. `file` unix utility for forensic analysis. In Oliver Goebel, Sandra Frings, Detlef Guenther, Jens Nedon, and Dirk Schadt, editors, *Proceedings of the IT Incident Management and IT Forensics (IMF'08)*, LNI140, pages 73–85. GI, September 2008.
  - [MLB07] Serguei A. Mokhov, Marc-André Laverdière, and Djamel Benredjem. Taxonomy of linux kernel vulnerability solutions. In *Innovative Techniques in Instruction Technology, E-learning, E-assessment, and Education*, pages 485–493, University of Bridgeport, U.S.A., 2007. Proceedings of CISSE/SCSS'07.
  - [Mok07] Serguei A. Mokhov. Introducing MARF: a modular audio recognition framework and its applications for scientific and software engineering research. In *Advances in Computer and Information Sciences and Engineering*, pages 473–478, University of Bridgeport, U.S.A., December 2007. Springer Netherlands. Proceedings of CISSE/SCSS'07.
  - [Mok08a] Serguei A. Mokhov. Encoding forensic multimedia evidence from MARF applications as Forensic Lucid expressions. In Tarek Sobh, Khaled Elleithy, and Ausif Mahmood, editors, *Novel Algorithms and Techniques in Telecommunications and Networking, proceedings of CISSE'08*, pages 413–416, University of Bridgeport, CT, USA, December 2008. Springer. Printed in January 2010.
  - [Mok08b] Serguei A. Mokhov. Study of best algorithm combinations for speech processing tasks in machine learning using median vs. mean clusters in MARF. In Bipin C. Desai, editor, *Proceedings of C3S2E'08*, pages 29–43, Montreal, Quebec, Canada, May 2008. ACM.
  - [Mok10a] Serguei A. Mokhov. Complete complimentary results report of the MARF's NLP approach to the DEFT 2010 competition. [online], June 2010. <http://arxiv.org/abs/1006.3787>.
  - [Mok10b] Serguei A. Mokhov. Evolution of MARF and its NLP framework. In *Proceedings of C3S2E'10*, pages 118–122. ACM, May 2010.
  - [Mok10c] Serguei A. Mokhov. L'approche MARF à DEFT 2010: A MARF approach to DEFT 2010. In *Proceedings of the 6th DEFT Workshop (DEFT'10)*, pages 35–49. LIMSI / ATALA, July 2010. DEFT 2010 Workshop at TALN 2010; online at [http://deft.limsi.fr/actes/2010/pdf/2\\_clac.pdf](http://deft.limsi.fr/actes/2010/pdf/2_clac.pdf).
  - [Mok10d] Serguei A. Mokhov. The use of machine learning with signal- and NLP processing of source code to fingerprint, detect, and classify vulnerabilities and weaknesses with MARFCAT. [online], October 2010. Online at <http://arxiv.org/abs/1010.2511>.
  - [Mok11] Serguei A. Mokhov. The use of machine learning with signal- and NLP processing of source code to fingerprint, detect, and classify vulnerabilities and weaknesses with MARFCAT. Technical Report NIST SP 500-283, NIST, October 2011. Report: [http://www.nist.gov/manuscript-publication-search.cfm?pub\\_id=909407](http://www.nist.gov/manuscript-publication-search.cfm?pub_id=909407), online e-print at <http://arxiv.org/abs/1010.2511>.
  - [Mok12] Serguei A. Mokhov. MARFCAT – MARF-based Code Analysis Tool. Published electronically within the MARF project, <http://sourceforge.net/projects/marf/files/Applications/MARFCAT/>, 2010–2012. Last viewed April 2012.
  - [Mot09] Motorola. Efficient polyphase FIR resampler for `numpy`: Native C/C++ implementation of the function `upfirdn()`. [online], 2009. <http://code.google.com/p/upfirdn/source/browse/upfirdn>.
  - [MPD08] Serguei A. Mokhov, Joey Paquet, and Mourad Debbabi. Formally specifying operational semantics and language constructs of Forensic Lucid. In Oliver Göbel, Sandra Frings, Detlef Günther, Jens Nedon, and Dirk Schadt, editors, *Proceedings of the IT Incident Management and IT Forensics (IMF'08)*, LNI140, pages 197–216. GI, September 2008. Online at <http://subs.emis.de/LNI/Proceedings/Proceedings140/gi-proc-140-014.pdf>.

- [MPD10] Serguei A. Mokhov, Joey Paquet, and Mourad Debbabi. Towards automatic deduction and event reconstruction using Forensic Lucid and probabilities to encode the IDS evidence. In S. Jha, R. Sommer, and C. Kreibich, editors, *Proceedings of RAID'10*, LNCS 6307, pages 508–509. Springer, September 2010.
- [MS02] Christopher D. Manning and Hinrich Schutze. *Foundations of Statistical Natural Language Processing*. MIT Press, 2002.
- [MSS09] Serguei A. Mokhov, Miao Song, and Ching Y. Suen. Writer identification using inexpensive signal processing techniques. In Tarek Sobh and Khaled Elleithy, editors, *Innovations in Computing Sciences and Software Engineering; Proceedings of CISSE'09*, pages 437–441. Springer, December 2009. ISBN: 978-90-481-9111-6, online at: <http://arxiv.org/abs/0912.5502>.
- [NIS12a] NIST. National Vulnerability Database. [online], 2005–2012. <http://nvd.nist.gov/>.
- [NIS12b] NIST. National Vulnerability Database statistics. [online], 2005–2012. <http://web.nvd.nist.gov/view/vuln/statistics>.
- [NJG<sup>+</sup>10] Vinod P. Nair, Harshit Jain, Yashwant K. Golecha, Manoj Singh Gaur, and Vijay Laxmi. MEDUSA: METamorphic malware dynamic analysis using signature from API. In *Proceedings of the 3rd International Conference on Security of Information and Networks*, SIN'10, pages 263–269, New York, NY, USA, 2010. ACM.
- [ODBN10] Vadim Okun, Aurelien Delaitre, Paul E. Black, and NIST SAMATE. Static Analysis Tool Exposition (SATE) 2010. [online], 2010. See <http://samate.nist.gov/SATE2010Workshop.html>.
- [ODBN12] Vadim Okun, Aurelien Delaitre, Paul E. Black, and NIST SAMATE. Static Analysis Tool Exposition (SATE) IV. [online], March 2012. See <http://samate.nist.gov/SATE.html>.
- [Paq09] Joey Paquet. Distributed educative execution of hybrid intensional programs. In *Proceedings of the 33rd Annual IEEE International Computer Software and Applications Conference (COMPSAC'09)*, pages 218–224, Seattle, Washington, USA, July 2009. IEEE Computer Society.
- [RM08] A. Newaz M. E. Rafiq and Yida Mao. A novel approach for automatic adjudication of new malware. In Nagib Callaos, William Lesso, C. Dale Zinn, Jorge Baralt, Jaouad Boukachour, Christopher White, Thilidzi Marwala, and Fulufhelo V. Nelwamondo, editors, *Proceedings of the 12th World Multi-Conference on Systemics, Cybernetics and Informatics (WM-SCI'08)*, volume V, pages 137–142, Orlando, Florida, USA, June 2008. IIIS.
- [Sch07] Rob Schreiber. MATLAB. *Scholarpedia*, 2(6):2929, 2007. <http://www.scholarpedia.org/article/MATLAB>.
- [SCL<sup>+</sup>03] Ivan Selesnick, Shihua Cai, Keyong Li, Levent Sendur, and A. Farras Abdelnour. MATLAB implementation of wavelet transforms. Technical report, Electrical Engineering, Polytechnic University, Brooklyn, NY, 2003. Online at <http://taco.poly.edu/WaveletSoftware/>.
- [SEZS01] M. G. Schultz, E. Eskin, E. Zadok, and S. J. Stolfo. Data mining methods for detection of new malicious executables. In *Proceedings of IEEE Symposium on Security and Privacy*, pages 38–49, Oakland, 2001.
- [Son10a] Dawn Song. BitBlaze: Security via binary analysis. [online], 2010. Online at <http://bitblaze.cs.berkeley.edu>.
- [Son10b] Dawn Song. WebBlaze: New techniques and tools for web security. [online], 2010. Online at <http://webblaze.cs.berkeley.edu>.
- [Sou12] Sourcefire. Snort: Open-source network intrusion prevention and detection system (IDS/IPS). [online], 1999–2012. <http://www.snort.org/>.
- [Sue07] M. Suenaga. Virus linguistics – searching for ethnic words. In H. Martin, editor, *Proceedings of the 17th Virus Bulletin International Conference*, pages 9–17, Vienna, Austria: ThePentagon, Abingdon, OX143YP, England, September 2007.
- [SXC<sup>+</sup>04] A. H. Sung, J. Xu, P. Chavez, , and S. Mukkamala. Static analyzer of vicious executables (SAVE). In *Proceedings of 20th Annual of Computer Security Applications Conference*, pages 326–334, December 2004.
- [The12] The MARF Research and Development Group. The Modular Audio Recognition Framework

and its Applications. [online], 2002–2012. <http://marf.sf.net> and <http://arxiv.org/abs/0905.1235>, last viewed April 2012.

- [Tli09] Syrine Tlili. *Automatic detection of safety and security vulnerabilities in open source software*. PhD thesis, Concordia Institute for Information Systems Engineering, Concordia University, Montreal, Canada, 2009. ISBN: 9780494634165.
- [Vas05] Emil Iordanov Vassev. General architecture for demand migration in the GIPSY demand-driven execution engine. Master's thesis, Department of Computer Science and Software Engineering, Concordia University, Montreal, Canada, June 2005. ISBN 0494102969.
- [VM12] Various contributors and MITRE. Common Weakness Enumeration (CWE) – a community-developed dictionary of software weakness types. [online], 2006–2012. See <http://cwe.mitre.org>.

## A Classification Result Tables

What follows are result tables with top classification results ranked from most precise at the top. This include the configuration settings for MARF by the means of options (the algorithm implementations are at their defaults [Mok07]).

## B Forensic Lucid Report Example

An example report encoding the reported data in Forensic Lucid for Wireshark 1.2.0 after using simple FFT-based feature extraction and Chebyshev distance as a classifier. The report provides the same data, compressed, as the SATE XML, but in the Forensic Lucid syntax for automated reasoning and event reconstruction during a digital investigation. The example is a an evidential statement context encoded for the use in the investigator's knowledge base of a particular case.

```
#FORENSICLUCID
evidential statement report_marfcats_0_0_2_SATE_IV_4
{
  weakness_1 @ [id:1, tool_specific_id:1, cweid:999, cwenam:"Insufficient Information (NVD-CWE-noinfo)"]
  where
    dimension id, tool_specific_id, cweid, cwenam;
    observation sequence weakness_1 = (locations_wk_1, 1, 0, 1.0);
    locations_wk_1 = locations @ [tool_specific_id:1, cweid:999, cwenam:"Insufficient Information (NVD-CWE-noinfo)"];
    observation location_id_340( [line => 828, path => "wireshark-1.2.0/epan/dissectors/packet-afs.c")
    observation location_id_340( [line => 1718, path => "wireshark-1.2.0/epan/dissectors/packet-afs.c")
    observation location_id_340( [line => 1729, path => "wireshark-1.2.0/epan/dissectors/packet-afs.c")
    observation location_id_340( [line => 1740, path => "wireshark-1.2.0/epan/dissectors/packet-afs.c")
    observation location_id_340( [line => 1747, path => "wireshark-1.2.0/epan/dissectors/packet-afs.c")
    textoutput="";
    observation grade = ([ severity => 5, tool_specific_rank => 0.0], 1, 0, 1.0);
  end;
  weakness_2 @ [id:2, tool_specific_id:2, cweid:119, cwenam:"Buffer Errors (CWE119)"]
  where
    dimension id, tool_specific_id, cweid, cwenam;
    observation sequence weakness_2 = (locations_wk_2, 1, 0, 1.0);
    locations_wk_2 = locations @ [tool_specific_id:2, cweid:119, cwenam:"Buffer Errors (CWE119)"];
    observation location_id_411( [line => 830, path => "wireshark-1.2.0/epan/dissectors/packet-ber.c")
    observation location_id_411( [line => 861, path => "wireshark-1.2.0/epan/dissectors/packet-ber.c")
    observation location_id_411( [line => 885, path => "wireshark-1.2.0/epan/dissectors/packet-ber.c")
    textoutput="";
    observation grade = ([ severity => 1, tool_specific_rank => 0.0], 1, 0, 1.0);
  end;
  weakness_3 @ [id:3, tool_specific_id:3, cweid:999, cwenam:"Insufficient Information (NVD-CWE-noinfo)"]
  where
    dimension id, tool_specific_id, cweid, cwenam;
    observation sequence weakness_3 = (locations_wk_3, 1, 0, 0.004878625561362933);
    locations_wk_3 = locations @ [tool_specific_id:3, cweid:999, cwenam:"Insufficient Information (NVD-CWE-noinfo)"];
    observation location_id_433( [line => 669, path => "wireshark-1.2.0/epan/dissectors/packet-btl2cap.c")
    textoutput="";
    observation grade = ([ severity => 5, tool_specific_rank => 204.97576364943077], 1, 0, 0.004878625561362933);
  end;
  weakness_4 @ [id:4, tool_specific_id:4, cweid:999, cwenam:"Insufficient Information (NVD-CWE-noinfo)"]
  where
    dimension id, tool_specific_id, cweid, cwenam;
    observation sequence weakness_4 = (locations_wk_4, 1, 0, 1.0);
    locations_wk_4 = locations @ [tool_specific_id:4, cweid:999, cwenam:"Insufficient Information (NVD-CWE-noinfo)"];
    observation location_id_550( [line => 248, path => "wireshark-1.2.0/epan/dissectors/packet-dcerpc-nt.c")
    observation location_id_550( [line => 252, path => "wireshark-1.2.0/epan/dissectors/packet-dcerpc-nt.c")
  end;
}
```

Table 1: CVE Stats for Wireshark 1.2.0, Separating DWT Wavelet Filter Preprocessing

guess	run	algorithms	good	bad	%
1st	1	-nopreprep -sdwt -fft -diff -spectrogram -graph -flucid	37	4	90.24
1st	2	-nopreprep -sdwt -fft -cheb -spectrogram -graph -flucid	37	4	90.24
1st	3	-nopreprep -sdwt -fft -eucl -spectrogram -graph -flucid	27	14	65.85
1st	4	-nopreprep -sdwt -fft -hamming -spectrogram -graph -flucid	26	15	63.41
1st	5	-nopreprep -sdwt -fft -mink -spectrogram -graph -flucid	22	19	53.66
1st	6	-nopreprep -sdwt -fft -cos -spectrogram -graph -flucid	38	65	36.89
2nd	1	-nopreprep -sdwt -fft -diff -spectrogram -graph -flucid	39	2	95.12
2nd	2	-nopreprep -sdwt -fft -cheb -spectrogram -graph -flucid	39	2	95.12
2nd	3	-nopreprep -sdwt -fft -eucl -spectrogram -graph -flucid	35	6	85.37
2nd	4	-nopreprep -sdwt -fft -hamming -spectrogram -graph -flucid	29	12	70.73
2nd	5	-nopreprep -sdwt -fft -mink -spectrogram -graph -flucid	31	10	75.61
2nd	6	-nopreprep -sdwt -fft -cos -spectrogram -graph -flucid	39	64	37.86
guess	run	class	good	bad	%
1st	1	CVE-2009-3829	6	0	100.00
1st	2	CVE-2009-2562	6	0	100.00
1st	3	CVE-2009-4378	6	0	100.00
1st	4	CVE-2010-2286	6	0	100.00
1st	5	CVE-2010-0304	6	0	100.00
1st	6	CVE-2009-4376	6	0	100.00
1st	7	CVE-2010-2283	6	0	100.00
1st	8	CVE-2009-3551	6	0	100.00
1st	9	CVE-2009-3550	6	0	100.00
1st	10	CVE-2009-3549	6	0	100.00
1st	11	CVE-2009-2563	6	2	75.00
1st	12	CVE-2009-2560	11	4	73.33
1st	13	CVE-2009-3241	15	9	62.50
1st	14	CVE-2010-1455	31	23	57.41
1st	15	CVE-2009-2561	6	6	50.00
1st	16	CVE-2010-2287	6	6	50.00
1st	17	CVE-2009-2559	6	6	50.00
1st	18	CVE-2009-3243	16	16	50.00
1st	19	CVE-2010-2285	6	7	46.15
1st	20	CVE-2009-4377	12	16	42.86
1st	21	CVE-2010-2284	6	9	40.00
1st	22	CVE-2009-3242	6	17	26.09
2nd	1	CVE-2009-3829	6	0	100.00
2nd	2	CVE-2009-2562	6	0	100.00
2nd	3	CVE-2009-4378	6	0	100.00
2nd	4	CVE-2010-2286	6	0	100.00
2nd	5	CVE-2010-0304	6	0	100.00
2nd	6	CVE-2009-4376	6	0	100.00
2nd	7	CVE-2010-2283	6	0	100.00
2nd	8	CVE-2009-3551	6	0	100.00
2nd	9	CVE-2009-3550	6	0	100.00
2nd	10	CVE-2009-3549	6	0	100.00
2nd	11	CVE-2009-2563	6	2	75.00
2nd	12	CVE-2009-2560	12	3	80.00
2nd	13	CVE-2009-3241	16	8	66.67
2nd	14	CVE-2010-1455	43	11	79.63
2nd	15	CVE-2009-2561	6	6	50.00
2nd	16	CVE-2010-2287	12	0	100.00
2nd	17	CVE-2009-2559	6	6	50.00
2nd	18	CVE-2009-3243	19	13	59.38
2nd	19	CVE-2010-2285	6	7	46.15
2nd	20	CVE-2009-4377	12	16	42.86
2nd	21	CVE-2010-2284	6	9	40.00
2nd	22	CVE-2009-3242	8	15	34.78

Table 2: CVE Stats for Wireshark 1.2.0, Low-Pass FFT Filter Preprocessing

guess	run	algorithms	good	bad	%
1st	1	-nopreprep -low -fft -cheb -flucid	37	4	90.24
1st	2	-nopreprep -low -fft -diff -flucid	37	4	90.24
1st	3	-nopreprep -low -fft -eucl -flucid	27	14	65.85
1st	4	-nopreprep -low -fft -hamming -flucid	23	18	56.10
1st	5	-nopreprep -low -fft -mink -flucid	22	19	53.66
1st	6	-nopreprep -low -fft -cos -flucid	36	114	24.00
2nd	1	-nopreprep -low -fft -cheb -flucid	38	3	92.68
2nd	2	-nopreprep -low -fft -diff -flucid	38	3	92.68
2nd	3	-nopreprep -low -fft -eucl -flucid	34	7	82.93
2nd	4	-nopreprep -low -fft -hamming -flucid	26	15	63.41
2nd	5	-nopreprep -low -fft -mink -flucid	31	10	75.61
2nd	6	-nopreprep -low -fft -cos -flucid	39	111	26.00
guess	run	class	good	bad	%
1st	1	CVE-2009-3829	6	0	100.00
1st	2	CVE-2009-4376	6	0	100.00
1st	3	CVE-2010-0304	6	0	100.00
1st	4	CVE-2010-2286	6	0	100.00
1st	5	CVE-2010-2283	6	0	100.00
1st	6	CVE-2009-3551	6	0	100.00
1st	7	CVE-2009-3549	6	0	100.00
1st	8	CVE-2009-3241	15	9	62.50
1st	9	CVE-2009-2560	9	6	60.00
1st	10	CVE-2010-1455	30	24	55.56
1st	11	CVE-2009-2563	6	5	54.55
1st	12	CVE-2009-2562	6	5	54.55
1st	13	CVE-2009-2561	6	7	46.15
1st	14	CVE-2009-4378	6	7	46.15
1st	15	CVE-2010-2287	6	7	46.15
1st	16	CVE-2009-3550	6	8	42.86
1st	17	CVE-2009-3243	13	23	36.11
1st	18	CVE-2009-4377	12	22	35.29
1st	19	CVE-2010-2285	6	11	35.29
1st	20	CVE-2009-2559	6	11	35.29
1st	21	CVE-2010-2284	6	12	33.33
1st	22	CVE-2009-3242	7	16	30.43
2nd	1	CVE-2009-3829	6	0	100.00
2nd	2	CVE-2009-4376	6	0	100.00
2nd	3	CVE-2010-0304	6	0	100.00
2nd	4	CVE-2010-2286	6	0	100.00
2nd	5	CVE-2010-2283	6	0	100.00
2nd	6	CVE-2009-3551	6	0	100.00
2nd	7	CVE-2009-3549	6	0	100.00
2nd	8	CVE-2009-3241	16	8	66.67
2nd	9	CVE-2009-2560	10	5	66.67
2nd	10	CVE-2010-1455	44	10	81.48
2nd	11	CVE-2009-2563	6	5	54.55
2nd	12	CVE-2009-2562	6	5	54.55
2nd	13	CVE-2009-2561	6	7	46.15
2nd	14	CVE-2009-4378	6	7	46.15
2nd	15	CVE-2010-2287	13	0	100.00
2nd	16	CVE-2009-3550	6	8	42.86
2nd	17	CVE-2009-3243	13	23	36.11
2nd	18	CVE-2009-4377	12	22	35.29
2nd	19	CVE-2010-2285	6	11	35.29
2nd	20	CVE-2009-2559	6	11	35.29
2nd	21	CVE-2010-2284	6	12	33.33
2nd	22	CVE-2009-3242	8	15	34.78

Table 3: CWE Stats for Wireshark 1.2.0, Separating DWT Wavelet Filter Preprocessing

guess	run	algorithms	good	bad	%
1st	1	-cweid -nopreprep -sdwt -fft -diff -flucid	31	5	86.11
1st	2	-cweid -nopreprep -sdwt -fft -eucl -flucid	29	7	80.56
1st	3	-cweid -nopreprep -sdwt -fft -mink -flucid	17	19	47.22
1st	4	-cweid -nopreprep -sdwt -fft -hamming -flucid	14	22	38.89
2nd	1	-cweid -nopreprep -sdwt -fft -diff -flucid	33	3	91.67
2nd	2	-cweid -nopreprep -sdwt -fft -eucl -flucid	34	2	94.44
2nd	3	-cweid -nopreprep -sdwt -fft -mink -flucid	27	9	75.00
2nd	4	-cweid -nopreprep -sdwt -fft -hamming -flucid	23	13	63.89
guess	run	class	good	bad	%
1st	1	CWE399	4	0	100.00
1st	2	CWE189	4	0	100.00
1st	3	NVD-CWE-Other	11	1	91.67
1st	4	CWE20	30	10	75.00
1st	5	NVD-CWE-noinfo	34	34	50.00
1st	6	CWE119	8	8	50.00
2nd	1	CWE399	4	0	100.00
2nd	2	CWE189	4	0	100.00
2nd	3	NVD-CWE-Other	11	1	91.67
2nd	4	CWE20	34	6	85.00
2nd	5	NVD-CWE-noinfo	53	15	77.94
2nd	6	CWE119	11	5	68.75

Table 4: CWE Stats for Wireshark 1.2.0, Low-Pass FFT Filter Preprocessing

guess	run	algorithms	good	bad	%
1st	1	-cweid -nopreprep -low -fft -diff -flucid	30	6	83.33
1st	2	-cweid -nopreprep -low -fft -cheb -flucid	30	6	83.33
1st	3	-cweid -nopreprep -low -fft -eucl -flucid	25	11	69.44
1st	4	-cweid -nopreprep -low -fft -mink -flucid	20	16	55.56
1st	5	-cweid -nopreprep -low -fft -cos -flucid	36	40	47.37
1st	6	-cweid -nopreprep -low -fft -hamming -flucid	12	24	33.33
2nd	1	-cweid -nopreprep -low -fft -diff -flucid	31	5	86.11
2nd	2	-cweid -nopreprep -low -fft -cheb -flucid	31	5	86.11
2nd	3	-cweid -nopreprep -low -fft -eucl -flucid	30	6	83.33
2nd	4	-cweid -nopreprep -low -fft -mink -flucid	22	14	61.11
2nd	5	-cweid -nopreprep -low -fft -cos -flucid	48	28	63.16
2nd	6	-cweid -nopreprep -low -fft -hamming -flucid	16	20	44.44
guess	run	class	good	bad	%
1st	1	CWE399	6	1	85.71
1st	2	CWE20	48	12	80.00
1st	3	NVD-CWE-Other	18	7	72.00
1st	4	CWE189	6	3	66.67
1st	5	NVD-CWE-noinfo	61	61	50.00
1st	6	CWE119	14	19	42.42
2nd	1	CWE399	6	1	85.71
2nd	2	CWE20	48	12	80.00
2nd	3	NVD-CWE-Other	18	7	72.00
2nd	4	CWE189	6	3	66.67
2nd	5	NVD-CWE-noinfo	78	44	63.93
2nd	6	CWE119	22	11	66.67

```

    observation location_id_550( [line => 256, path => "wireshark-1.2.0/epan/dissectors/packet-dcerpc-nt.c")
    observation location_id_550( [line => 1138, path => "wireshark-1.2.0/epan/dissectors/packet-dcerpc-nt.c")
    observation location_id_550( [line => 1142, path => "wireshark-1.2.0/epan/dissectors/packet-dcerpc-nt.c")
    observation location_id_550( [line => 1146, path => "wireshark-1.2.0/epan/dissectors/packet-dcerpc-nt.c")
    observation location_id_550( [line => 1201, path => "wireshark-1.2.0/epan/dissectors/packet-dcerpc-nt.c")
    observation location_id_550( [line => 1205, path => "wireshark-1.2.0/epan/dissectors/packet-dcerpc-nt.c")
    observation location_id_550( [line => 1209, path => "wireshark-1.2.0/epan/dissectors/packet-dcerpc-nt.c")
    observation location_id_550( [line => 1314, path => "wireshark-1.2.0/epan/dissectors/packet-dcerpc-nt.c")
    observation location_id_550( [line => 1318, path => "wireshark-1.2.0/epan/dissectors/packet-dcerpc-nt.c")
    observation location_id_550( [line => 1322, path => "wireshark-1.2.0/epan/dissectors/packet-dcerpc-nt.c")
    textoutput="";
    observation grade = ([ severity => 5, tool_specific_rank => 0.0], 1, 0, 1.0);
end;
weakness_5 @ [id:5, tool_specific_id:5, cweid:999, cwenam:"Insufficient Information (NVD-CWE-noinfo)"]
where
    dimension id, tool_specific_id, cweid, cwenam;
    observation sequence weakness_5 = (locations_wk_5, 1, 0, 0.003778693428627627);
    locations_wk_5 = locations @ [tool_specific_id:5, cweid:999, cwenam:"Insufficient Information (NVD-CWE-noinfo)"];
    observation location_id_552( [line => 77, path => "wireshark-1.2.0/epan/dissectors/packet-dtls.c")
    textoutput="";
    observation grade = ([ severity => 5, tool_specific_rank => 264.64173897356557], 1, 0, 0.003778693428627627);
end;
weakness_6 @ [id:6, tool_specific_id:6, cweid:999, cwenam:"Insufficient Information (NVD-CWE-noinfo)"]
where
    dimension id, tool_specific_id, cweid, cwenam;
    observation sequence weakness_6 = (locations_wk_6, 1, 0, 0.004125022212036806);
    locations_wk_6 = locations @ [tool_specific_id:6, cweid:999, cwenam:"Insufficient Information (NVD-CWE-noinfo)"];
    observation location_id_763( [line => 8447, path => "wireshark-1.2.0/epan/dissectors/packet-gsm_a_rr.c")
    textoutput="";
    observation grade = ([ severity => 5, tool_specific_rank => 242.42293704067873], 1, 0, 0.004125022212036806);
end;
weakness_7 @ [id:7, tool_specific_id:7, cweid:999, cwenam:"Insufficient Information (NVD-CWE-noinfo)"]
where
    dimension id, tool_specific_id, cweid, cwenam;
    observation sequence weakness_7 = (locations_wk_7, 1, 0, 1.0);
    locations_wk_7 = locations @ [tool_specific_id:7, cweid:999, cwenam:"Insufficient Information (NVD-CWE-noinfo)"];
    observation location_id_863( [line => 945, path => "wireshark-1.2.0/epan/dissectors/packet-infiniband.c")
    textoutput="";
    observation grade = ([ severity => 5, tool_specific_rank => 0.0], 1, 0, 1.0);
end;
weakness_8 @ [id:8, tool_specific_id:8, cweid:119, cwenam:"Buffer Errors (CWE119)"]
where
    dimension id, tool_specific_id, cweid, cwenam;
    observation sequence weakness_8 = (locations_wk_8, 1, 0, 1.0);
    locations_wk_8 = locations @ [tool_specific_id:8, cweid:119, cwenam:"Buffer Errors (CWE119)"];
    observation location_id_877( [line => 2746, path => "wireshark-1.2.0/epan/dissectors/packet-ipmi-se.c")
    observation location_id_877( [line => 2748, path => "wireshark-1.2.0/epan/dissectors/packet-ipmi-se.c")
    observation location_id_877( [line => 2752, path => "wireshark-1.2.0/epan/dissectors/packet-ipmi-se.c")
    textoutput="";
    observation grade = ([ severity => 1, tool_specific_rank => 0.0], 1, 0, 1.0);
end;
weakness_9 @ [id:9, tool_specific_id:9, cweid:998, cwenam:"Other (NVD-CWE-Other)"]
where
    dimension id, tool_specific_id, cweid, cwenam;
    observation sequence weakness_9 = (locations_wk_9, 1, 0, 1.0);
    locations_wk_9 = locations @ [tool_specific_id:9, cweid:998, cwenam:"Other (NVD-CWE-Other)"];
    observation location_id_882( [line => 792, path => "wireshark-1.2.0/epan/dissectors/packet-ipmi.c")
    textoutput="";
    observation grade = ([ severity => 5, tool_specific_rank => 0.0], 1, 0, 1.0);
end;
weakness_10 @ [id:10, tool_specific_id:10, cweid:119, cwenam:"Buffer Errors (CWE119)"]
where
    dimension id, tool_specific_id, cweid, cwenam;
    observation sequence weakness_10 = (locations_wk_10, 1, 0, 1.0);
    locations_wk_10 = locations @ [tool_specific_id:10, cweid:119, cwenam:"Buffer Errors (CWE119)"];
    observation location_id_969( [line => 523, path => "wireshark-1.2.0/epan/dissectors/packet-lwres.c")
    textoutput="";
    observation grade = ([ severity => 1, tool_specific_rank => 0.0], 1, 0, 1.0);
end;
weakness_11 @ [id:11, tool_specific_id:11, cweid:20, cwenam:"Input Validation (CWE20)"]
where
    dimension id, tool_specific_id, cweid, cwenam;
    observation sequence weakness_11 = (locations_wk_11, 1, 0, 1.0);
    locations_wk_11 = locations @ [tool_specific_id:11, cweid:20, cwenam:"Input Validation (CWE20)"];
    observation location_id_1099( [line => 62, path => "wireshark-1.2.0/epan/dissectors/packet-paltalk.c")
    textoutput="";
    observation grade = ([ severity => 1, tool_specific_rank => 0.0], 1, 0, 1.0);
end;
weakness_12 @ [id:12, tool_specific_id:12, cweid:999, cwenam:"Insufficient Information (NVD-CWE-noinfo)"]
where
    dimension id, tool_specific_id, cweid, cwenam;
    observation sequence weakness_12 = (locations_wk_12, 1, 0, 0.004878625561362927);
    locations_wk_12 = locations @ [tool_specific_id:12, cweid:999, cwenam:"Insufficient Information (NVD-CWE-noinfo)"];
    observation location_id_1174( [line => 897, path => "wireshark-1.2.0/epan/dissectors/packet-radius.c")
    observation location_id_1174( [line => 906, path => "wireshark-1.2.0/epan/dissectors/packet-radius.c")
    observation location_id_1174( [line => 913, path => "wireshark-1.2.0/epan/dissectors/packet-radius.c")
    observation location_id_1174( [line => 1005, path => "wireshark-1.2.0/epan/dissectors/packet-radius.c")
    observation location_id_1174( [line => 1227, path => "wireshark-1.2.0/epan/dissectors/packet-radius.c")
    textoutput="";
    observation grade = ([ severity => 5, tool_specific_rank => 204.975763649431], 1, 0, 0.004878625561362927);
end;
weakness_13 @ [id:13, tool_specific_id:13, cweid:999, cwenam:"Insufficient Information (NVD-CWE-noinfo)"]

```



```

where
  dimension id, tool_specific_id, cweid, cwename;
  observation sequence weakness_13 = (locations_wk_13, 1, 0, 1.0);
  locations_wk_13 = locations @ [tool_specific_id:13, cweid:999, cwename:"Insufficient Information (NVD-CWE-noinfo)"];
  observation location_id_1282( [line => 1131, path => "wireshark-1.2.0/epan/dissectors/packet-sflow.c"]
  textoutput="";
  observation grade = ([ severity => 5, tool_specific_rank => 0.0], 1, 0, 1.0);
end;
weakness_14 @ [id:14, tool_specific_id:14, cweid:998, cwename:"Other (NVD-CWE-Other)"]
where
  dimension id, tool_specific_id, cweid, cwename;
  observation sequence weakness_14 = (locations_wk_14, 1, 0, 1.0);
  locations_wk_14 = locations @ [tool_specific_id:14, cweid:998, cwename:"Other (NVD-CWE-Other)"];
  observation location_id_1303( [line => 2141, path => "wireshark-1.2.0/epan/dissectors/packet-smb-pipe.c"]
  textoutput="";
  observation grade = ([ severity => 5, tool_specific_rank => 0.0], 1, 0, 1.0);
end;
weakness_15 @ [id:15, tool_specific_id:15, cweid:998, cwename:"Other (NVD-CWE-Other)"]
where
  dimension id, tool_specific_id, cweid, cwename;
  observation sequence weakness_15 = (locations_wk_15, 1, 0, 1.0);
  locations_wk_15 = locations @ [tool_specific_id:15, cweid:998, cwename:"Other (NVD-CWE-Other)"];
  observation location_id_1307( [line => 8757, path => "wireshark-1.2.0/epan/dissectors/packet-smb.c"]
  textoutput="";
  observation grade = ([ severity => 5, tool_specific_rank => 0.0], 1, 0, 1.0);
end;
weakness_16 @ [id:16, tool_specific_id:16, cweid:189, cwename:"Numeric Errors (CWE189)"]
where
  dimension id, tool_specific_id, cweid, cwename;
  observation sequence weakness_16 = (locations_wk_16, 1, 0, 1.0);
  locations_wk_16 = locations @ [tool_specific_id:16, cweid:189, cwename:"Numeric Errors (CWE189)"];
  observation location_id_1307( [line => 2195, path => "wireshark-1.2.0/epan/dissectors/packet-smb.c"]
  textoutput="";
  observation grade = ([ severity => 2, tool_specific_rank => 0.0], 1, 0, 1.0);
end;
weakness_17 @ [id:17, tool_specific_id:17, cweid:999, cwename:"Insufficient Information (NVD-CWE-noinfo)"]
where
  dimension id, tool_specific_id, cweid, cwename;
  observation sequence weakness_17 = (locations_wk_17, 1, 0, 0.008328136212759968);
  locations_wk_17 = locations @ [tool_specific_id:17, cweid:999, cwename:"Insufficient Information (NVD-CWE-noinfo)"];
  observation location_id_1307( [line => 8457, path => "wireshark-1.2.0/epan/dissectors/packet-smb.c"]
  textoutput="";
  observation grade = ([ severity => 5, tool_specific_rank => 120.07488523877028], 1, 0, 0.008328136212759968);
end;
weakness_18 @ [id:18, tool_specific_id:18, cweid:999, cwename:"Insufficient Information (NVD-CWE-noinfo)"]
where
  dimension id, tool_specific_id, cweid, cwename;
  observation sequence weakness_18 = (locations_wk_18, 1, 0, 0.008328136212759964);
  locations_wk_18 = locations @ [tool_specific_id:18, cweid:999, cwename:"Insufficient Information (NVD-CWE-noinfo)"];
  observation location_id_1309( [line => 955, path => "wireshark-1.2.0/epan/dissectors/packet-smb2.c"]
  textoutput="";
  observation grade = ([ severity => 5, tool_specific_rank => 120.07488523877032], 1, 0, 0.008328136212759964);
end;
weakness_19 @ [id:19, tool_specific_id:19, cweid:999, cwename:"Insufficient Information (NVD-CWE-noinfo)"]
where
  dimension id, tool_specific_id, cweid, cwename;
  observation sequence weakness_19 = (locations_wk_19, 1, 0, 0.004321352067642762);
  locations_wk_19 = locations @ [tool_specific_id:19, cweid:999, cwename:"Insufficient Information (NVD-CWE-noinfo)"];
  observation location_id_1333( [line => 813, path => "wireshark-1.2.0/epan/dissectors/packet-ssl-utils.c"]
  observation location_id_1333( [line => 843, path => "wireshark-1.2.0/epan/dissectors/packet-ssl-utils.c"]
  textoutput="";
  observation grade = ([ severity => 5, tool_specific_rank => 231.40905539443497], 1, 0, 0.004321352067642762);
end;
weakness_20 @ [id:20, tool_specific_id:20, cweid:999, cwename:"Insufficient Information (NVD-CWE-noinfo)"]
where
  dimension id, tool_specific_id, cweid, cwename;
  observation sequence weakness_20 = (locations_wk_20, 1, 0, 0.0021114804997331383);
  locations_wk_20 = locations @ [tool_specific_id:20, cweid:999, cwename:"Insufficient Information (NVD-CWE-noinfo)"];
  observation location_id_1334( [line => 153, path => "wireshark-1.2.0/epan/dissectors/packet-ssl-utils.h"]
  textoutput="";
  observation grade = ([ severity => 5, tool_specific_rank => 473.60134281438354], 1, 0, 0.0021114804997331383);
end;
weakness_21 @ [id:21, tool_specific_id:21, cweid:999, cwename:"Insufficient Information (NVD-CWE-noinfo)"]
where
  dimension id, tool_specific_id, cweid, cwename;
  observation sequence weakness_21 = (locations_wk_21, 1, 0, 0.003463630817441021);
  locations_wk_21 = locations @ [tool_specific_id:21, cweid:999, cwename:"Insufficient Information (NVD-CWE-noinfo)"];
  observation location_id_1335( [line => 275, path => "wireshark-1.2.0/epan/dissectors/packet-ssl.c"]
  textoutput="";
  observation grade = ([ severity => 5, tool_specific_rank => 288.71437306901373], 1, 0, 0.003463630817441021);
end;
weakness_22 @ [id:22, tool_specific_id:22, cweid:999, cwename:"Insufficient Information (NVD-CWE-noinfo)"]
where
  dimension id, tool_specific_id, cweid, cwename;
  observation sequence weakness_22 = (locations_wk_22, 1, 0, 0.00412502212036806);
  locations_wk_22 = locations @ [tool_specific_id:22, cweid:999, cwename:"Insufficient Information (NVD-CWE-noinfo)"];
  observation location_id_1583( [line => 1799, path => "wireshark-1.2.0/epan/packet.c"]
  textoutput="";
  observation grade = ([ severity => 5, tool_specific_rank => 242.42293704067873], 1, 0, 0.00412502212036806);
end;
weakness_23 @ [id:23, tool_specific_id:23, cweid:399, cwename:"Resource Management Errors (CWE399)"]
where

```

```

dimension id, tool_specific_id, cweid, cwename;
observation sequence weakness_23 = (locations_wk_23, 1, 0, 1.0);
locations_wk_23 = locations @ [tool_specific_id:23, cweid:399, cwename:"Resource Management Errors (CWE399)"];
  observation location_id_1611( [line => 345, path => "wireshark-1.2.0/epan/sigcomp-udvm.c")
  textoutput="";
  observation grade = ([ severity => 3, tool_specific_rank => 0.0], 1, 0, 1.0);
end;
weakness_24 @ [id:24, tool_specific_id:24, cweid:119, cwename:"Buffer Errors (CWE119)"]
where
  dimension id, tool_specific_id, cweid, cwename;
  observation sequence weakness_24 = (locations_wk_24, 1, 0, 1.0);
  locations_wk_24 = locations @ [tool_specific_id:24, cweid:119, cwename:"Buffer Errors (CWE119)"];
    observation location_id_1611( [line => 321, path => "wireshark-1.2.0/epan/sigcomp-udvm.c")
    textoutput="";
    observation grade = ([ severity => 1, tool_specific_rank => 0.0], 1, 0, 1.0);
end;
weakness_25 @ [id:25, tool_specific_id:25, cweid:20, cwename:"Input Validation (CWE20)"]
where
  dimension id, tool_specific_id, cweid, cwename;
  observation sequence weakness_25 = (locations_wk_25, 1, 0, 0.001495003320843141);
  locations_wk_25 = locations @ [tool_specific_id:25, cweid:20, cwename:"Input Validation (CWE20)"];
    observation location_id_2012( [line => 89, path => "wireshark-1.2.0/plugins/docsis/packet-bpkmreq.c")
    textoutput="";
    observation grade = ([ severity => 1, tool_specific_rank => 668.8948352542972], 1, 0, 0.001495003320843141);
end;
weakness_26 @ [id:26, tool_specific_id:26, cweid:20, cwename:"Input Validation (CWE20)"]
where
  dimension id, tool_specific_id, cweid, cwename;
  observation sequence weakness_26 = (locations_wk_26, 1, 0, 0.0014959114047375394);
  locations_wk_26 = locations @ [tool_specific_id:26, cweid:20, cwename:"Input Validation (CWE20)"];
    observation location_id_2013( [line => 90, path => "wireshark-1.2.0/plugins/docsis/packet-bpkmrsp.c")
    textoutput="";
    observation grade = ([ severity => 1, tool_specific_rank => 668.4887867242726], 1, 0, 0.0014959114047375394);
end;
weakness_27 @ [id:27, tool_specific_id:27, cweid:20, cwename:"Input Validation (CWE20)"]
where
  dimension id, tool_specific_id, cweid, cwename;
  observation sequence weakness_27 = (locations_wk_27, 1, 0, 0.002153585613826869);
  locations_wk_27 = locations @ [tool_specific_id:27, cweid:20, cwename:"Input Validation (CWE20)"];
    observation location_id_2020( [line => 72, path => "wireshark-1.2.0/plugins/docsis/packet-dsaack.c")
    textoutput="";
    observation grade = ([ severity => 1, tool_specific_rank => 464.341883405798], 1, 0, 0.002153585613826869);
end;
weakness_28 @ [id:28, tool_specific_id:28, cweid:20, cwename:"Input Validation (CWE20)"]
where
  dimension id, tool_specific_id, cweid, cwename;
  observation sequence weakness_28 = (locations_wk_28, 1, 0, 0.00229165238295895);
  locations_wk_28 = locations @ [tool_specific_id:28, cweid:20, cwename:"Input Validation (CWE20)"];
    observation location_id_2022( [line => 72, path => "wireshark-1.2.0/plugins/docsis/packet-dsarsp.c")
    textoutput="";
    observation grade = ([ severity => 1, tool_specific_rank => 436.36635618741343], 1, 0, 0.00229165238295895);
end;
weakness_29 @ [id:29, tool_specific_id:29, cweid:20, cwename:"Input Validation (CWE20)"]
where
  dimension id, tool_specific_id, cweid, cwename;
  observation sequence weakness_29 = (locations_wk_29, 1, 0, 0.002184230355798278);
  locations_wk_29 = locations @ [tool_specific_id:29, cweid:20, cwename:"Input Validation (CWE20)"];
    observation location_id_2023( [line => 72, path => "wireshark-1.2.0/plugins/docsis/packet-dscack.c")
    textoutput="";
    observation grade = ([ severity => 1, tool_specific_rank => 457.82716889058463], 1, 0, 0.002184230355798278);
end;
weakness_30 @ [id:30, tool_specific_id:30, cweid:20, cwename:"Input Validation (CWE20)"]
where
  dimension id, tool_specific_id, cweid, cwename;
  observation sequence weakness_30 = (locations_wk_30, 1, 0, 0.0023006295251237975);
  locations_wk_30 = locations @ [tool_specific_id:30, cweid:20, cwename:"Input Validation (CWE20)"];
    observation location_id_2025( [line => 73, path => "wireshark-1.2.0/plugins/docsis/packet-dscrsp.c")
    textoutput="";
    observation grade = ([ severity => 1, tool_specific_rank => 434.6636383996635], 1, 0, 0.0023006295251237975);
end;
weakness_31 @ [id:31, tool_specific_id:31, cweid:20, cwename:"Input Validation (CWE20)"]
where
  dimension id, tool_specific_id, cweid, cwename;
  observation sequence weakness_31 = (locations_wk_31, 1, 0, 0.001897888480826156);
  locations_wk_31 = locations @ [tool_specific_id:31, cweid:20, cwename:"Input Validation (CWE20)"];
    observation location_id_2032( [line => 72, path => "wireshark-1.2.0/plugins/docsis/packet-regack.c")
    textoutput="";
    observation grade = ([ severity => 1, tool_specific_rank => 526.9013485790784], 1, 0, 0.001897888480826156);
end;
weakness_32 @ [id:32, tool_specific_id:32, cweid:20, cwename:"Input Validation (CWE20)"]
where
  dimension id, tool_specific_id, cweid, cwename;
  observation sequence weakness_32 = (locations_wk_32, 1, 0, 0.002216818195096963);
  locations_wk_32 = locations @ [tool_specific_id:32, cweid:20, cwename:"Input Validation (CWE20)"];
    observation location_id_2035( [line => 73, path => "wireshark-1.2.0/plugins/docsis/packet-regrsp.c")
    textoutput="";
    observation grade = ([ severity => 1, tool_specific_rank => 451.096983149879], 1, 0, 0.002216818195096963);
end;
weakness_33 @ [id:33, tool_specific_id:33, cweid:999, cwename:"Insufficient Information (NVD-CWE-noinfo)"]
where
  dimension id, tool_specific_id, cweid, cwename;
  observation sequence weakness_33 = (locations_wk_33, 1, 0, 0.0028814675905206645);

```

```

locations_wk_33 = locations @ [tool_specific_id:33, cweid:999, cwename:"Insufficient Information (NVD-CWE-noinfo)"];
observation location_id_2097( [line => 433, path => "wireshark-1.2.0/plugins/pcua/pcua_complextypeparser.c")
textoutput="";
observation grade = ([ severity => 5, tool_specific_rank => 347.04537482557834], 1, 0, 0.0028814675905206645);
end;
weakness_34 @ [id:34, tool_specific_id:34, cweid:999, cwename:"Insufficient Information (NVD-CWE-noinfo)"]
where
dimension id, tool_specific_id, cweid, cwename;
observation sequence weakness_34 = (locations_wk_34, 1, 0, 0.0028288900371324934);
locations_wk_34 = locations @ [tool_specific_id:34, cweid:999, cwename:"Insufficient Information (NVD-CWE-noinfo)"];
observation location_id_2107( [line => 616, path => "wireshark-1.2.0/plugins/pcua/pcua_serviceparser.c")
textoutput="";
observation grade = ([ severity => 5, tool_specific_rank => 353.49553601371184], 1, 0, 0.0028288900371324934);
end;
weakness_35 @ [id:35, tool_specific_id:35, cweid:999, cwename:"Insufficient Information (NVD-CWE-noinfo)"]
where
dimension id, tool_specific_id, cweid, cwename;
observation sequence weakness_35 = (locations_wk_35, 1, 0, 0.003058220966230374);
locations_wk_35 = locations @ [tool_specific_id:35, cweid:999, cwename:"Insufficient Information (NVD-CWE-noinfo)"];
observation location_id_2110( [line => 340, path => "wireshark-1.2.0/plugins/pcua/pcua_simpletypes.c")
textoutput="";
observation grade = ([ severity => 5, tool_specific_rank => 326.9874907805045], 1, 0, 0.003058220966230374);
end;
weakness_36 @ [id:36, tool_specific_id:36, cweid:999, cwename:"Insufficient Information (NVD-CWE-noinfo)"]
where
dimension id, tool_specific_id, cweid, cwename;
observation sequence weakness_36 = (locations_wk_36, 1, 0, 0.0018096494904338023);
locations_wk_36 = locations @ [tool_specific_id:36, cweid:999, cwename:"Insufficient Information (NVD-CWE-noinfo)"];
observation location_id_2112( [line => 132, path => "wireshark-1.2.0/plugins/pcua/pcua_transport_layer.c")
observation location_id_2112( [line => 169, path => "wireshark-1.2.0/plugins/pcua/pcua_transport_layer.c")
observation location_id_2112( [line => 181, path => "wireshark-1.2.0/plugins/pcua/pcua_transport_layer.c")
observation location_id_2112( [line => 195, path => "wireshark-1.2.0/plugins/pcua/pcua_transport_layer.c")
observation location_id_2112( [line => 226, path => "wireshark-1.2.0/plugins/pcua/pcua_transport_layer.c")
observation location_id_2112( [line => 250, path => "wireshark-1.2.0/plugins/pcua/pcua_transport_layer.c")
textoutput="";
observation grade = ([ severity => 5, tool_specific_rank => 552.593198454295], 1, 0, 0.0018096494904338023);
end;
weakness_37 @ [id:37, tool_specific_id:37, cweid:119, cwename:"Buffer Errors (CWE119)"]
where
dimension id, tool_specific_id, cweid, cwename;
observation sequence weakness_37 = (locations_wk_37, 1, 0, 1.0);
locations_wk_37 = locations @ [tool_specific_id:37, cweid:119, cwename:"Buffer Errors (CWE119)"];
observation location_id_2321( [line => 149, path => "wireshark-1.2.0/wiretap/daintree-sna.c")
observation location_id_2321( [line => 205, path => "wireshark-1.2.0/wiretap/daintree-sna.c")
textoutput="";
observation grade = ([ severity => 1, tool_specific_rank => 0.0], 1, 0, 1.0);
end;
weakness_38 @ [id:38, tool_specific_id:38, cweid:189, cwename:"Numeric Errors (CWE189)"]
where
dimension id, tool_specific_id, cweid, cwename;
observation sequence weakness_38 = (locations_wk_38, 1, 0, 1.0);
locations_wk_38 = locations @ [tool_specific_id:38, cweid:189, cwename:"Numeric Errors (CWE189)"];
observation location_id_2327( [line => 228, path => "wireshark-1.2.0/wiretap/erf.c")
textoutput="";
observation grade = ([ severity => 2, tool_specific_rank => 0.0], 1, 0, 1.0);
end;
}

```

# Index

## API

- DEFT2010App, 2
- FormItem, 9
- ResultSet, 9
- Warning, 8
- WriterIdentApp, 2

C, 3–7, 10, 21

C++, 3, 7, 21

## Chrome

- 5.0.375.54, 3, 10, 12, 15, 17
- 5.0.375.70, 3

## CVE

- CVE-2009-2559, 24, 25
- CVE-2009-2560, 24, 25
- CVE-2009-2561, 24, 25
- CVE-2009-2562, 11–13, 20, 24, 25
- CVE-2009-2563, 24, 25
- CVE-2009-3241, 24, 25
- CVE-2009-3242, 24, 25
- CVE-2009-3243, 24, 25
- CVE-2009-3549, 24, 25
- CVE-2009-3550, 24, 25
- CVE-2009-3551, 24, 25
- CVE-2009-3829, 24, 25
- CVE-2009-4376, 24, 25
- CVE-2009-4377, 24, 25
- CVE-2009-4378, 24, 25
- CVE-2010-0304, 24, 25
- CVE-2010-1455, 24, 25
- CVE-2010-2283, 24, 25
- CVE-2010-2284, 24, 25
- CVE-2010-2285, 24, 25
- CVE-2010-2286, 24, 25
- CVE-2010-2287, 24, 25

## CWE

- CWE-119, 14–17
- CWE-16, 17
- CWE-189, 14, 16
- CWE-20, 14–17
- CWE-200, 15, 17
- CWE-22, 15, 17
- CWE-255, 15, 17
- CWE-264, 15, 17
- CWE-399, 16, 17
- CWE-79, 15–17

CWE-94, 16, 17  
NVD-CWE-noinfo, 14, 16, 17, 26  
NVD-CWE-Other, 14, 16, 17, 26

Dovecot 1.2.0, 3  
Dovecot 1.2.17, 3  
Dovecot 1.2.x, 10  
Dovecot 2.0.beta6.20100626, 4

#### Files

\*\_test.xml, 18, 19  
\*\_train.xml, 18, 19  
collect-files-meta-synthetic.pl, 5  
collect-files-meta.pl, 5  
packet-afs.c, 11–13, 20  
report-cweidnopreprepcharunigramadddelta-train-test-test-run-quick-tomcat-5-5-33-cwe-nlp.xml, 19  
report-cweidnoprepreprawfftcheb-wireshark-1.2.0-half-train-cwe.xml, 13  
report-cweidnoprepreprawfftcos-train-test-test-run-quick-tomcat-5-5-33-cwe.xml, 18  
report-cweidnoprepreprawfftdiff-wireshark-1.2.0-half-train-cwe.xml, 13  
report-nopreprepcharunigramadddelta-train-test-test-run-quick-tomcat-5-5-33-cve-nlp.xml, 18  
report-noprepreprawfftcheb-train-test-test-run-quick-tomcat-5-5-33-cve.xml, 18  
report-noprepreprawfftcheb-train-test-test-run-quick-wireshark-1-2-18-cve.xml, 19  
report-noprepreprawfftcos-train-test-test-run-quick-tomcat-5-5-33-cve.xml, 18  
report-noprepreprawfftcos-train-test-test-run-quick-wireshark-1-2-18-cve.xml, 19  
report-noprepreprawfftdiff-train-test-test-run-quick-wireshark-1-2-18-cve.xml, 19  
report-noprepreprawffteucl-train-test-test-run-quick-wireshark-1-2-18-cve.xml, 19  
report-noprepreprawffthamming-train-test-test-run-quick-wireshark-1-2-18-cve.xml, 19  
report-noprepreprawfftmink-train-test-test-run-quick-wireshark-1-2-18-cve.xml, 19

Forensic Lucid, 9, 21–23

#### Frameworks

MARF, 1, 2, 7, 8, 22, 23

GIPSY, 7, 8, 22

Java, 3, 4, 6, 7, 10, 21

#### Jetty

6.1.16, 3  
6.1.26, 3  
6.1.x, 10

Jini, 7

JMS, 7

#### Libraries

MARF, 1, 2, 7, 8, 22, 23

MARF, 1, 2, 7, 8, 22, 23

#### Applications

MARFCAT, 1, 2, 4, 5, 7–10, 22

MARFCAT, 1, 2, 4, 5, 7–10, 22

#### Options

-char, 16, 17

-cheb, 14, 16, 24–26

-cos, 14, 16, 18, 19, 24–26

-cweid, 14–17, 26

-diff, 14, 15, 24–26

-eucl, 14, 16, 24–26

-fft, 14–16, 24–26

-flucid, 24–26

-graph, 24

-hamming, 14–16, 24–26

-low, 25, 26

-mink, 14, 16, 19, 24–26

-nopreprep, 14–17, 24–26

-raw, 14–16, 20

-sdwt, 24, 26

-spectrogram, 24

-unigram, 16, 17

Pebble, 4, 7, 8

Perl, 5

PHP, 3

#### Test cases

Chrome 5.0.375.54, 3, 10, 12, 15, 17

Chrome 5.0.375.70, 3

Dovecot 1.2.0, 3

Dovecot 1.2.17, 3

Dovecot 1.2.x, 10

Dovecot 2.0.beta6.20100626, 4

Jetty 6.1.16, 3

Jetty 6.1.26, 3

Jetty 6.1.x, 10

Pebble 2.5-M2, 4, 7, 8

Tomcat 5.5.13, 3, 7, 8, 12, 14, 17–19

Tomcat 5.5.29, 3, 18

Tomcat 5.5.33, 3, 7, 8, 10, 18

Wireshark 1.2.0, 3, 11–13, 16, 18–20, 23–26

Wireshark 1.2.18, 3, 10, 12, 13, 18, 19

- Wireshark 1.2.9, 3, 12, 13, 18
- Wordpress 2.0, 3
- Wordpress 2.2.3, 3
- Wordpress 2.x, 10
- TODO, 19
- Tomcat
  - 5.5.13, 3, 7, 8, 12, 14, 17–19
  - 5.5.29, 3, 18
  - 5.5.33, 3, 7, 8, 10, 18
- Tools
  - codegen, 7
- Wireshark
  - 1.2.0, 3, 11–13, 16, 18–20, 23–26
  - 1.2.18, 3, 10, 12, 13, 18, 19
  - 1.2.9, 3, 12, 13, 18
- Wordpress
  - 2.0, 3
  - 2.2.3, 3
  - 2.x, 10